Doctoral School of the Universitat Jaume I
Doctoral Programme in Computer Science

DEFINITION AND ANALYSIS OF STRATEGIC PREDICTIVE INDICATORS FOR SOCIAL NETWORKS

Ph. D. dissertation
Indira Lázara Lanza Cruz

Supervisor
Dr. Rafael Berlanga Llavori

Castellón de La Plana, November, 2023

Doctoral Programme in Computer Science
Doctoral School of the Universitat Jaume I

DEFINITION AND ANALYSIS OF STRATEGIC PREDICTIVE INDICATORS FOR SOCIAL NETWORKS

A dissertation submited by Indira Lázara Lanza Cruz to obtain the degree of Doctor of Philosophy by the Universitat Jaume I

Author
Indira Lázara Lanza Cruz

Supervisor
Dr. Rafael Berlanga Llavori

INDIRA LAZARA|LANZA|CRUZ
Firmado digitalmente por INDIRA LAZARA|LANZA|CRUZ
Fecha: 2023.11.20 10:23:15 +01'00'

RAFAEL|BERLANGA|LLAVORI
Firmado digitalmente por RAFAEL|BERLANGA|LLAVORI
Fecha: 2023.11.20 10:44:09 +01'00'

Castellón de La Plana, November, 2023

This dissertation is funded by:

The pre-doctoral grant of the Universitat Jaume I with reference PREDOC/2017/28

OCIT Research staff mobility program with reference E-2019-16

Spanish Ministry of Industry and Commerce grant number TIN2017-88805-R

Spanish Ministry of Industry and Commerce grant number PDC2021-121097-I00

Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature

Thesis by compendium of the following publications:

Indexed Journals:

- Lanza Cruz, Indira, Berlanga, Rafael and Aramburu, María José. (2023). Multidimensional Author Profiling for Social Business Intelligence. *Information Systems Frontiers*. Springer. https://doi.org/10.1007/s10796-023-10370-0. (Q1)
- Lanza Cruz, Indira; Berlanga, Rafael and Aramburu, María José. (2018). Modeling Analytical Streams for Social Business Intelligence. *Informatics*, 5, 33. https://doi.org/10.3390/informatics5030033. (Q1)
- Aramburu, María José; Berlanga, Rafael; Lanza Cruz, Indira. (2020). Social Media Multidimensional Analysis for Intelligent Health Surveillance. *International Journal of Environmental Research and Public Health*, 17, 2289. https://doi.org/10.3390/ijerph17072289. (Q2)
- Aramburu, María José; Berlanga, Rafael and Lanza Cruz, Indira. (2023). A Data Quality Multidimensional Model for Social Media Analysis. *Business & Information Systems Engineering*, DOI: 10.1007/s12599-023-00840-9 (In Press). (Q1)

International Conferences:

- Lanza Cruz, Indira and Berlanga Llavori, Rafael. (2018). Defining Dynamic Indicators for Social Network Analysis: A Case Study in the Automotive Domain using Twitter. *In Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR*, ISBN 978-989-758-330-8; ISSN 2184-3228, pages 221-228. DOI: 10.5220/0006932902210228.
- Berlanga, R. ; Lanza Cruz, Indira and Aramburu, María José. (2019). Quality Indicators for Social Business Intelligence. *In 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Granada, Spain, 2019, pp. 229-236, doi: 10.1109/SNAMS.2019.8931862.

This thesis has the acceptance of the co-authors of the publications that the doctoral student presents as a thesis and their express waiver to present them as part of another doctoral thesis.

Other scientific contributions derived from this research, which are not presented as chapters in this thesis:

- Lanza Cruz, Indira; Aramburu Cabo, María José and Berlanga, Rafael. (2016). Metodología de inteligencia de negocio para análisis social en la infraestructura de datos enlazados SLOD-BI. Ciencia da informacao. Vol 45, No.3, paginas 199-215. ISSN: 0100-1965. (Q4)
- Aramburu, M.; Berlanga, R. and Lanza Cruz, Indira. (2021). Quality Management in Social Business Intelligence Projects. *In Proceedings of the 23rd International Conference on Enterprise Information Systems - Volume 1: ICEIS*, ISBN 978-989-758-509-8; ISSN 2184-4992, pages 320-327. DOI: 10.5220/0010495703200327.
- Berlanga, Rafael; Aramburu, María José; Lanza Cruz, Indira ; Llidó Escrivá, Dolores María; Museros, Lledó and Sanz, Ismael.(2018). Dynamic SLOD-BI: Infraestructura Dinámica de Inteligencia de Negocio Social. *Sánchez-Figueroa, F. (Ed.), Actas de las XXIII Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2018)*. Sevilla, septiembre de 2018. Publisher SISTEDES.
- Berlanga, Rafael; Jimeno Yepes, Antonio; Pérez, María and Lanza Cruz, Indira. (2018). Coarse-grained Semantic Characterization of Large Knowledge Resources. *In Proceedings of the 5th Spanish Conference on Information Retrieval (CERI '18)*. Association for Computing Machinery, New York, NY, USA, Article 16, 1–4. https://doi.org/10.1145/3230599.3230616.
- Soriano, Mario; Berlanga, Rafael and Lanza Cruz, Indira. (2023). On the problem of automatically aligning indicators to SDGs. *In: Pesquita, C., et al. The Semantic Web: ESWC 2023 Satellite Events. ESWC 2023. Lecture Notes in Computer Science*, vol 13998. Springer, Cham. https://doi.org/10.1007/978-3-031-43458-7_26.

Agradecimientos

En primer lugar, me complace expresar mi profundo agradecimiento a mi supervisor de tesis Rafael Berlanga Llavori por su invaluable apoyo y guía durante la realización de mi tesis doctoral. Agradezco su constante orientación, sus conocimientos expertos y dedicación a lo largo de este desafiante proceso de investigación. Su compromiso y mentoría han sido fundamentales en mi desarrollo académico contribuyendo de manera significativa a la calidad y el éxito de mi tesis. Estoy agradecida por su influencia en mi desarrollo profesional y por la confianza que depositó en mí.

También deseo expresar mi gratitud a la profesora María José Aramburu Cabo, supervisora y compañera de investigación. Valoro enormemente su disponibilidad y disposición para brindar retroalimentación constructiva, orientación metodológica y recomendaciones valiosas que han fortalecido mi trabajo de investigación.

Asimismo, deseo expresar mi gratitud a mis compañeros de trabajo por su apoyo, intercambio de ideas y motivación a lo largo de este recorrido académico. Sus contribuciones y colaboración han enriquecido mi experiencia y han sido fundamentales para superar los desafíos que se presentaron en el camino.

También quiero agradecer a los miembros del comité de expertos que revisaron mi tesis por su valioso tiempo, comentarios y sugerencias que han enriquecido y mejorado mi trabajo. Sus aportes críticos y perspicaces han sido esenciales para mi crecimiento como investigadora.

En especial, agradecer el apoyo incondicional de mi familia y seres queridos. En primer lugar, agradezco a mi esposo, por su inquebrantable apoyo, paciencia y amor durante toda esta larga etapa de estudios. Su aliento constante ha sido mi fuente de inspiración y fortaleza.

Agradezco eternamente a mis padres por su amor y por ser mis fuentes de inspiración en la vida. A mi madre por ser mi modelo a seguir como mujer y como profesional. A mi padre por su cariño, apoyo y optimismo para seguir adelante. Agradezco profundamente a mi hermana por siempre estar a mi lado, por acompañarme a caminar por este recorrido a veces complejo pero muy hermoso.

En este momento culminante de mi trayectoria académica, quiero dedicar un sincero agradecimiento a mí misma por la perseverancia, dedicación y valentía que he demostrado a lo largo de este arduo viaje hacia la obtención de mi doctorado. Este logro no solo representa el dominio de un campo de conocimiento, sino también la superación de desafíos personales y la capacidad de enfrentar y aprender de las adversidades.

Resumen

Las redes sociales se han convertido en poderosos canales de comunicación global que conectan a personas de todas partes del mundo a través de intereses comunes. Plataformas populares como Facebook, Twitter y LinkedIn han revolucionado la forma en que interactuamos y han dejado una huella significativa en nuestros hábitos de vida. Estos medios sociales no solo han facilitado la interacción entre individuos, sino que también han generado un enorme interés en el ámbito académico y empresarial. La información compartida públicamente en estas redes y los patrones de interacción entre usuarios han despertado un gran interés en la investigación del comportamiento humano en diversos campos, como las ciencias sociales, la economía y el marketing. Los datos generados en las redes sociales brindan una valiosa oportunidad para comprender mejor las preferencias, tendencias y necesidades de las personas en una escala masiva. En este sentido, las empresas están dedicando considerables esfuerzos para analizar el contenido de las redes sociales y obtener información relevante sobre sus clientes. Sin embargo, se enfrentan a desafíos significativos al intentar adoptar técnicas y metodologías que permitan un análisis más efectivo. La gran cantidad de datos, la complejidad de los algoritmos y las constantes actualizaciones en las plataformas sociales hacen que el análisis de las redes sociales sea un campo en constante evolución que requiere adaptabilidad y conocimientos especializados.

Tradicionalmente la disciplina de inteligencia de negocio (BI, por sus siglas en inglés, Business Intelligence) se ha limitado a definir indicadores estratégicos a partir de datos temporales recogidos internamente en la organización. Sin embargo, existe una gran cantidad de información estratégica que puede influir en el comportamiento de una empresa y se encuentra en fuentes externas, especialmente en los medios sociales. Es por esta razón que surge la Inteligencia de Negocio Social (SBI, por sus siglas en inglés, Social Business Intelligence) como una convergencia innovadora entre dos áreas de investigación distintas: la BI y las redes sociales.

La nueva disciplina SBI, se centra en la captura y análisis de información relevante y actual difundida en las redes sociales, con el propósito de respaldar la toma de decisiones empresariales. A pesar de su importancia y potencial, esta disciplina enfrenta diversos desafíos y problemas de investigación que requieren ser superados. En primer lugar, la complejidad, heterogeneidad y los niveles de calidad de los datos dificultan su procesamiento y análisis efectivo. Además, la interpretación de los datos puede ser complicada debido a la naturaleza subjetiva de las opiniones y actitudes sociales. Estos factores pueden generar errores en las decisiones empresariales, por lo que es fundamental desarrollar técnicas de análisis efectivas para abordar este problema. Otro desafío importante es lograr la integración efectiva de datos provenientes de diversas fuentes, manteniendo siempre en consideración la privacidad y ética en el uso de dichos datos con fines comerciales. Asimismo, la falta de flexibilidad en los modelos existentes los hace inadecuados para abordar las necesidades analíticas dinámicas que surgen en este campo. Además, los obstáculos relacionados con los datos, sistemas y tecnologías propietarias añaden dificultades adicionales a la implementación y avance de la disciplina. Todo esto dificulta la implementación de soluciones integrales que aborden la definición de

indicadores estratégicos, los cuales son fundamentales para medir el progreso hacia la consecución de objetivos en los medios sociales y apoyar el proceso de toma decisiones.

Abordar el análisis de redes sociales desde un punto de vista metodológico, tal y como se ha hecho en la disciplina de BI con herramientas como los almacenes de datos, implica enfrentarse a peculiaridades propias del dominio, así como a la complejidad del tratamiento de grandes volúmenes de datos (big data). Esta investigación presenta desafíos adicionales, entre ellos el modelado de indicadores que deben ajustarse a las cambiantes necesidades de análisis y a datos volátiles, cuyo comportamiento es difícil de predecir. Además, se hace necesaria la creación de infraestructuras de datos abiertos que faciliten la integración con datos externos, con el objetivo de promover la colaboración y la innovación en los distintos sectores de negocio.

Esta tesis se centra en una nueva línea de investigación dentro del área de SBI que tiene como objetivo los indicadores sociales. En este contexto, el estudio se centra en proponer nuevas metodologías generales para descubrir y describir indicadores sociales de forma semiautomática a partir de métricas sociales y contextos de análisis dinámicos.

La principal contribución de esta tesis es la propuesta de un marco metodológico integral para el desarrollo de proyectos de SBI. El mismo aborda una infraestructura semántica para la representación de patrones analíticos de datos relevantes para el análisis social, así como para la definición y seguimiento de indicadores sociales estratégicos. En este marco se abstraen varios procesos metodológicos que pueden ser adaptados y extendidos a diferentes tareas de análisis y dominios de negocio. Específicamente, se propone un formalismo multidimensional para representar y evaluar indicadores sociales a partir de hechos capturados en streaming derivados de los datos de redes sociales. El enfoque permite la definición de indicadores sociales bajo demanda, así como el tratamiento de dimensiones y métricas de forma dinámica. Además, se propone una arquitectura tipo Kappa para el procesamiento en streaming de indicadores sociales. La misma permite integrar tareas de Ciencia de Datos y Análisis de Datos en una misma área de trabajo. La arquitectura se basa tanto en el uso de datos vinculados como en modelos multidimensionales, lo que facilita el enriquecimiento semántico de los datos y les da forma según los requisitos de análisis.

Con el objetivo de caracterizar tanto a los creadores de contenido en las redes, como a su audiencia, se presenta un novedoso método no-supervisado para la clasificación de autores en redes sociales aplicando técnicas de análisis multidimensional y clasificación automática de textos. Para la tarea de clasificación se definen etiquetas multidimensionales basadas en roles de negocio que se pueden ajustar de forma dinámica. El método arroja resultados competitivos al ser evaluado en diversos conjuntos de datos, incluyendo aquellos obtenidos del estado del arte. El perfilado multidimensional de autores puede ayudar a los sistemas de información a caracterizar la audiencia de los temas y noticias populares publicados en redes sociales. El análisis de calidad de los datos también puede beneficiarse en gran medida del perfilado del autor, ya que es un indicador directo de la calidad de las fuentes de datos.

Por otra parte, se propone un método para crear colecciones de datos de alta calidad destinadas al análisis de las redes sociales. Este método hace uso de un novedoso modelo de datos multidimensional para la construcción de cubos con medidas de impacto para diversas métricas de calidad. Además, los cubos de calidad incluyen una dimensión usuario-rol para que las métricas de calidad puedan evaluarse en función de los roles de negocio de los usuarios. El método posibilita la generación automática de indicadores de calidad adaptados a las necesidades de análisis. La implicación práctica principal es que permite a los analistas medir la calidad de los datos procesados desde distintas perspectivas, considerando las categorías de los perfiles de los usuarios que generan los contenidos. Se

demuestra que analizar las métricas de calidad en función de las categorías de usuario facilita la comprensión de los datos y proporciona estrategias adecuadas para obtener perspectivas o conocimientos relevantes para el negocio.

Para demostrar la aplicabilidad de la infraestructura y los métodos propuestos en escenarios reales, se llevaron a cabo evaluaciones en diferentes casos de uso en los sectores de automoción, turismo sostenible y vigilancia inteligente en el ámbito de la salud. Durante el estudio relacionado con el análisis social para la vigilancia inteligente en salud, se integraron todos los métodos propuestos en la tesis. Los resultados obtenidos demuestran que el modelo dinámico multidimensional propuesto permite identificar eventos y temas relevantes para el dominio, así como analizar su audiencia e impacto de manera efectiva.

En resumen, el sistema desarrollado proporciona a las organizaciones una infraestructura para generar indicadores estratégicos a partir de los datos recuperados de las redes sociales, lo cual resulta sumamente útil para medir el impacto de sus acciones en los medios digitales. El marco metodológico ofrece mecanismos flexibles para enriquecer semánticamente los datos sociales y definir nuevos indicadores. Los modelos semánticos propuestos facilitan la integración de los resultados en la nube de datos, lo que permite establecer redes de colaboración entre organizaciones. Por último, la infraestructura propuesta facilita el descubrimiento de conocimiento a partir de la información generada en las redes sociales desde una perspectiva de SBI.

Abstract

Social media has become powerful channels of global communication that connect people from all around the world through common interests. Popular platforms like Facebook, Twitter, and LinkedIn have revolutionized the way we interact and have made a significant impact on our lifestyle habits. These social media platforms have not only facilitated interaction between individuals but have also generated immense interest in the academic and business domains. The publicly shared information on these networks and the interaction patterns among users have sparked great interest in researching human behaviour in various fields such as social sciences, economics, and marketing. The data generated on social media provides a valuable opportunity to gain a better understanding of people's preferences, trends, and needs on a massive scale. In this regard, companies are devoting considerable efforts to analyse social media content and obtain relevant information about their customers. However, they face significant challenges when trying to adopt techniques and methodologies that enable effective analysis. The large volume of data, complexity of algorithms, and constant platform updates make social media analysis a constantly evolving field that requires adaptability and specialized knowledge.

Traditionally, the discipline of Business Intelligence (BI) has been limited to defining strategic indicators based on internally collected temporal data within the organization. However, there is a wealth of strategic information that can influence a company's behaviour and is found in external sources, especially on social media. It is for this reason that Social Business Intelligence (SBI) emerges as an innovative convergence between two distinct areas of research: BI and social networks.

The emerging discipline of SBI (Social Business Intelligence) focuses on capturing and analysing relevant and up-to-date information disseminated on social media platforms, with the purpose of supporting business decision-making. Despite its importance and potential, this discipline faces various challenges and research problems that need to be overcome. Firstly, the complexity, heterogeneity, and quality levels of the data make their processing and analysis a challenging task. Additionally, the subjective nature of social opinions and attitudes can complicate data interpretation, leading to potential errors in business decisions. It is crucial, therefore, to develop effective analysis techniques to address these challenges. Another significant challenge lies in achieving effective integration of data from diverse sources while always considering privacy and ethical considerations in the use of such data for commercial purposes. Furthermore, the lack of flexibility in existing models renders them inadequate to address the dynamic analytical needs that arise in this field. Moreover, obstacles related to proprietary data, systems, and technologies add further difficulties to the implementation and advancement of the discipline. All these factors hinder the implementation of comprehensive solutions that address the definition of strategic indicators, which are vital for measuring progress towards achieving goals in social media and supporting the decision-making process.

Addressing social media analysis from a methodological standpoint, similar to how it has been done in the field of BI with tools like data warehouses, involves dealing with domain-specific peculiarities as well as the complexity of handling large volumes of data

(big data). This research presents additional challenges, including the modelling of indicators that need to adapt to changing analysis needs and volatile data, whose behaviour is difficult to predict. Furthermore, the creation of open data infrastructures is necessary to facilitate integration with external data, with the goal of promoting collaboration and innovation across different business sectors.

This thesis focuses on a new research direction within the field of SBI, specifically targeting social indicators. In this context, the study aims to propose new general methodologies for the semi-automatic discovery and description of social indicators based on social metrics and dynamic analysis contexts.

The main contribution of this thesis is the proposal of a comprehensive methodological framework for the development of SBI projects. It addresses a semantic infrastructure for representing analytical patterns of data relevant to social analysis, as well as the definition and tracking of strategic social indicators. Within this framework, several methodological processes are abstracted, which can be adapted and extended to different analysis tasks and business domains. Specifically, a multidimensional formalism is proposed to represent and evaluate social indicators based on streaming facts derived from social media data. This approach allows for the on-demand definition of social indicators, as well as the dynamic treatment of dimensions and metrics. Additionally, a Kappa-like architecture is proposed for streaming processing of social indicators, enabling the integration of Data Science and Data Analysis tasks in a unified workspace. The architecture is based on both linked data and multidimensional models, facilitating the semantic enrichment of data and shaping it according to analysis requirements.

In order to characterize both content creators on social media networks and their audience, a novel unsupervised method is presented for author classification on social networks, applying multidimensional analysis techniques and automatic text classification. For the classification task, multidimensional labels based on business roles are defined, which can be dynamically adjusted. The method yields competitive results when evaluated on various datasets, including those obtained from state-of-the-art sources. Multidimensional author profiling can assist information systems in characterizing the audience of popular topics and news published on social networks. The analysis of data quality can also greatly benefit from author profiling as it serves as a direct indicator of the quality of data sources.

Additionally, a method is proposed for creating high-quality data collections intended for social network analysis. This method employs a novel multidimensional data model for constructing cubes with impact measures for various quality metrics. Furthermore, the quality cubes include a user-role dimension to enable quality metrics evaluation based on users' business roles. The method enables the automatic generation of tailored quality indicators that align with analysis needs. The primary practical implication is that it allows analysts to measure the quality of processed data from different perspectives, considering user profile categories that generate the content. It is demonstrated that analysing quality metrics based on user categories facilitates data comprehension and provides appropriate strategies for obtaining relevant business insights.

To demonstrate the applicability of the proposed infrastructure and methods in real-world scenarios, evaluations were conducted in different use cases in the automotive, sustainable tourism, and intelligent surveillance in healthcare sectors. During the study related to social analysis for intelligent healthcare surveillance, all the proposed methods in the thesis were integrated. The obtained results demonstrate that the proposed dynamic multidimensional model allows for the identification of relevant events and topics in the domain, as well as effective analysis of their audience and impact.

In summary, the developed system provides organizations with an infrastructure to generate strategic indicators from data retrieved from social networks, which is highly useful for measuring the impact of their actions in digital media. The methodological framework offers flexible mechanisms for semantic enrichment of social data and defining new indicators. The proposed semantic models facilitate the integration of results in the data cloud, enabling collaboration networks among organizations. Finally, the proposed infrastructure facilitates knowledge discovery from information generated in social networks from an SBI perspective.

Contents

Chapter 1

Introduction

1.1 Motivation

Social networks, such as Facebook, Twitter, and LinkedIn, have fostered close connections among Internet users globally. These connections are built upon shared interests, spanning various aspects like politics, professions, religion, emotions, and recreational pursuits. On average, Internet users worldwide dedicate approximately two and a half hours per day to engaging with social platforms (Coyne, Rogers, Zurcher, Stockdale, & Booth, 2020). This innate human desire to communicate, coupled with the advancements of the Web of Data (Heath & Bizer, 2022), further enriches these information sources.

Social network analysis is widely used in the social and behavioural sciences, as well as in economics, marketing, and industrial engineering. The social network perspective focuses on the relationships between social entities; examples include communications between members of a group, economic transactions between corporations, and trade between nations (Knoke & Yang, 2019).

There is a great interest in the business world in analysing the content of social networks in order to understand customer needs and to respond proactively and efficiently to those needs (Dwivedi, Ismagilova, Rana, & Raman, 2023) (Parveen Tajudeen, Ismawati Jaafar, & Ainin, 2017). There are a large number of proposals in the literature related to social networks (Peng, Yu, & Mueller, Social networking big data: Opportunities, solutions, and challenges, 2018) and sentiment analysis (Zucco, Calabrese, Agapito, Guzzi, & Cannataro, 2020), to discover, respectively, patterns within the relationships between individuals and qualitative aspects in their conversations. However, companies still

struggle to adopt, implement, and standardize methodologies and techniques for an effective social network analysis tool (Ruhi, 2014).

The Business Intelligence (BI) discipline is devoted to defining strategic indicators based on measures of interest defined on a set of temporary data collected from different sources and integrated under the same multidimensional schema. Traditionally, the data collected comes from relational databases and limited external sources from other businesses or customers. However, much of the relevant strategic information that can affect an organization currently resides in external sources, mainly in social networks and media. Recent research on social network analysis has emphasized the need to adopt a BI-based approach to collecting, analysing, and interpreting social network data. To implement the above, it is necessary for each social metric to be linked to a Key Performance Indicator (KPI), which in turn should correspond to a strategic business objective. However, social media metrics are rarely combined with other business measures to calculate KPIs of different purpose (Ruhi, 2014).

The Social Business Model represents a new paradigm for businesses, involving the integration of social networks into their daily operations. These networks serve as an ideal platform for digital marketing, as well as to capture the Voice of the Market (VoM) and the Voice of the Customer (VoC) (Berlanga, et al., 2015). The social and economic influence of these platforms is expanding, and companies must not overlook them in their strategic plans if they wish to remain competitive. Within this context, a new approach to Business Intelligence (BI) emerges, shifting the focus from traditional service execution (such as sales and promotions) to the contents generated by individuals on social networks (such as preferences, opinions, and experiences). Consequently, the concept of Social Business Intelligence (SBI) has emerged as a novel discipline in both academic and business literature, resulting from the convergence of two distinct research domains: BI and social networks. The field of SBI involves the capture and analysis of pertinent and up-to-date information disseminated on social networks, assisting companies in making informed decisions. Many of the SBI approaches try to extend the results of BI models to the field of social networks (Berlanga, et al., 2015). However, the lack of flexibility of these models makes them unsuitable in scenarios with highly changing analytical demands. On the other hand, there are few approaches that address social analysis through the specification of indicators derived from strategic objectives (Maté, Trujillo, & Mylopoulos, 2017). Currently, the SBI domain is largely uncharted, marked by controversial definitions of terms and concepts, fragmented and isolated research efforts, and obstacles posed by proprietary data and immature technologies (Gioti, Ponis, & Panayiotou, 2018).

At present, the analysis of social networks has attained a level of maturity that warrants a more methodological approach, similar to how traditional BI approached through the utilization of data warehouses (Diamantini, Potena, & Storti, 2016) (Horkoff, Barone, & Jiang, 2014) (Maté, Trujillo, & Mylopoulos, 2012). However, there are peculiarities in this domain that do not allow a direct adaptation of traditional BI (Berlanga, et al., 2015): the social data to be analysed is considered big data, and the social indicators are dynamic, volatile and less predictable in their behaviour. The main challenges to be addressed can be categorized into the following points: (i) the reliability and quality of the external data sources, (ii) the nature of unstructured data, and (ii) the integration of business intelligence and social network analysis (Abu Salih, Wongthongtham, Beheshti, & Zajabbari, 2019).

In this sense, this thesis addresses a new research line in the field of SBI, with a specific focus on the development and application of social indicators as the primary objective. This research aims to contribute to the understanding and effective utilization of social indicators for evaluating and analysing social phenomena within the context of SBI. The

primary focus is to develop innovative approaches that streamline the process of identifying and characterizing social indicators by leveraging social metrics data and domain-specific topics. To transform the collected social network data into meaningful social indicators, novel automated methods are proposed to handle streaming data and employ text-mining techniques.

An indispensable requirement when approaching SBI projects is to verify the quality of the data to be processed. With this objective in mind, this thesis also addresses the study of methodologies to obtain high-quality social data collections, considering the data source and the analysis context. The methods for defining quality indicators were validated through multiple aspects, including the nature of the author's profile and considering various perspectives of quality, such as credibility, reputation, usefulness, and completeness.

1.2 Background

1.2.1 Business Intelligence on Social Networks

Business Intelligence (BI) is the utilization of data within a company to enhance decision-making processes. BI encompasses a wide array of methodologies, technologies, and processes for collecting, storing, accessing, and analysing data from various sources. This facilitates the extraction of profiles, segmentation, risk modelling, sales forecasting, and other predictive insights that support business decisions. Common BI functions include reporting, online analytical processing (OLAP), data mining, complex event processing, benchmarking, text mining, predictive analytics, and prescriptive analytics. Traditional BI architecture typically involves a data warehouse, an integration layer for data transformation, and an analysis layer for extracting valuable knowledge. The main goal of BI is to derive strategic knowledge, often in the form of Key Performance Indicators (KPIs) (Maté, Trujillo, & Mylopoulos, 2017) (Roldán-García, García-Nieto, Maté, Trujillo, & Aldana-Montes, 2021), from data collected within an organization. *Social Business Intelligence* has recently emerged as a fusion of BI and social media, expanding data sources and enhancing decision-making capabilities (Francia, Gallinucci, Golfarelli, & Rizzi, Social Business Intelligence in Action, 2016) (Francia, Gallinucci, & Golfarelli, Social BI to understand the debate on vaccines on the Web and social media: unraveling the anti-, free, and pro-vax communities, 2019) (Dwivedi, Ismagilova, Rana, & Raman, 2023). The main difference with traditional BI is that data sources have increased from transactional and limited external data sources to many other data sources such as data coming from global environment in forms of news, economic factors, social networks, web blogs, etc. In the context of social media analytics as part of a business intelligence initiative, the alignment of social media metrics with business metrics and KPIs is crucial, all of which should ultimately support strategic business objectives like revenue growth, cost reduction, and customer satisfaction. Figure 1 represents the concept of SBI, which merges various fields of research that have been extensively treated independently in the literature, such as: BI, SMA and Text Mining.

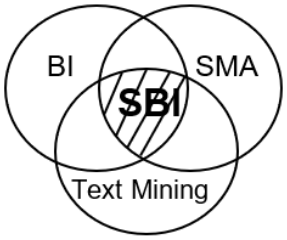


Figure 1 - Relationship between the research fields that Social Business Intelligence encompasses.

This thesis introduces the concept of a *social indicator* (Salvatore, Biffignandi, & Bianchi, 2021), a temporary measure enabling dynamic assessment of an organization's social network activities, all linked to the organization's social-oriented strategic objectives. The primary challenges in this research encompass dynamically defining effective social indicators, considering the vast volume of constantly published social network data, combining challenges from BI and big data management. The central challenge of effectively utilizing social data to model them as strategic social indicators and enabling interoperability capabilities for decision-making within companies will be addressed in this thesis.

1.2.2 Semantic Technologies for Social Network Analysis

Semantic models are abstractions that describe a situation in a particular way, hiding certain details while illuminating others. They are used to help people understand complex phenomena and to communicate, explain, and make predictions about the world. In the context of the Semantic Web, semantic modelling is the activity of distilling communal knowledge out of a chaotic mess of information. The Semantic Web standards have been created as a medium in which people can collaborate on models to organize the information that they share and advance the common collection of knowledge (Allemang & Hendler, 2011).

Semantic technologies are a set of tools and techniques used to create, store, and exchange semantic data on the Web. These technologies include semantic markup languages as RDF(Resource Description Framework), OWL (Web Ontology Language), SPARQL (SPARQL Protocol and RDF Query Language), and other Semantic Web languages and standards. These technologies facilitate the development of ontologies, serving as semantic models to represent knowledge within a specific domain. These ontologies provide a formal description of a set of concepts and their relationships, enabling the representation of knowledge in a machine-readable format. By using Semantic technologies, it is possible to create smarter applications that can integrate and analyse data from multiple sources, and to build more efficient and effective information systems. Semantic technologies for social network analysis focus on the meaning and interpretation of content, making them useful for understanding and extracting valuable information from online social networks.

This emphasis on semantics aligns with the principles of Web 3.0 and Linked Open Data (LOD) cloud, where data is published and linked through Unique Resource Identifiers (URIs) with well-defined semantics. RDF, a key component of this framework, offers a

standardized approach for expressing relationships and connecting diverse data on the web. RDF enhances the comprehension and interconnectedness of information, fostering a more comprehensive understanding for both users and machines (Berlanga, et al., 2015).

In the context of this thesis, the data infrastructure is designed using these technologies. Primarily, we implement analytical models through RDF-based knowledge graphs (Jiménez-Ruiz, Hassanzadeh, Efthymiou, Chen, & Srinivas, 2020) which are represented as a set of RDF triples <*s, p, o*>, where *"s"* represents a subject (a class or an instance), *"p"* represents a predicate (a property) and *"o"* represents an object (a class, an instance or a data value, e.g., text, date and number). Knowledge graphs are a way to connect and explain concepts/data in a flexible, accessible, and reusable manner. Knowledge graphs find application in various domains, including the semantic web, artificial intelligence, online search, and knowledge management, facilitating data comprehension, navigation, interoperability, and information enrichment. The combination of these technologies with social network analysis techniques enables researchers and professionals to extract deeper and more meaningful insights from online interactions, identify patterns, trends, communities, and make informed decisions in areas such as marketing, public opinion research, among other applications.

1.2.3 Quality Indicators and Author Profiling

Evaluating the quality of social data is a context-dependent process that requires the identification of the most suitable *quality indicators* for SBI projects (Salvatore, Biffignandi, & Bianchi, 2021). These indicators are metrics that assess the overall quality of a data collection by combining measures obtained by various quality criteria, aiding in the filtration of relevant messages for an SBI project. However, the selection of the best criteria and metrics and their integration into a quality indicator is a complex task that necessitates a deep understanding of the context and analysis objectives.

On the other hand, the use of authorship analysis techniques, such as *author profiling*, can be valuable for assessing the quality of content on social networks. Author profiling involves discerning between different author classes by examining the shared language patterns among individuals. This technique involves analysing a set of texts to uncover various author characteristics based on linguistic aspects. Common characteristics include age, gender, personality traits, and occupation (Daelemans, et al., 2019). The process typically includes identifying specific text features, creating a representation of the target profile (e.g., Bag-of-Words model), and building a classification model for the target profile. Author profiling is divided into two main approaches: a traditional one, effective for formal texts like books and the press (utilizing lexical and syntactical features), and another focused on informal documents like blogs and tweets (emphasizing content and stylistic features). Research into author profiling techniques on social networks is attracting increasing interest, as evidenced by the numerous competitions organized within the framework of PAN at CLEF (Conference and Labs of the Evaluation Forum that consist in a series of scientific events and shared tasks on digital text forensics and stylometry) and PAN at FIRE (Forum for Information Retrieval Evaluation) from 2010 to the present (Potthast, Rosso, Stamatatos, & Stein, 2019).

In the context of this thesis, we explore how author profiling can help determine the credibility of content authors on social networks by analysing their linguistic characteristics, contributing to the identification of low-quality authors and directing high-quality content to interested audiences.

1.3 Social Media BI approaches

Nowadays many processes related to decision-making in any organization are affected by social media trends as they offer immediate feedback from the VoC and the VoM. Social networks have become a fundamental part of the information ecosystem whose efficient analysis is of great value to consumers and companies. The analysis potential offered by social networks has been demonstrated in multiple studies in different domains. For example, related to the tasks of opinion mining and sentiment analysis, there are solutions for the analysis of communication, social behaviour, political sentiment and knowing the state of opinion about a brand or topic (Liu, et al., 2013) (Rosenthal, Farra, & Nakov, 2017) (Berlanga, et al., 2015). In the field of user profiling, it has been primarily applied to the determination of political polarization (Taulé, Rangel, Martí, & Rosso, 2018), demographic studies (such as age estimation, gender, ethnicity, language variety, etc.) (Daelemans, et al., 2019), (Potthast, Rosso, Stamatatos, & Stein, 2019), (Wood-Doughty, Mahajan, & Dredze, 2018), and identification of interests. The study of published posts has been used to identify emerging social debate (Francia, Gallinucci, & Golfarelli, 2019), detection of events in real time such as protests and demonstrations, diseases, catastrophes and population displacement. The study of the structure of a social network and the interactions between users to understand how information flows, understand social phenomena, identify influential users, among other applications.

Table 1 summarises the main analytical tasks and methods proposed in the academic literature for social analytics and business intelligence. In the literature reviewed, we identified a large number of analytical tasks, which are grouped into six categories, namely: Sentiment Analysis, Topic Modelling, Author Profiling, Predictive Modelling and Network-based Analysis.

Analytical approach	**Analysis tasks**	**Methods**	**Approaches**
Sentiment Analysis/ Opinion Mining	Detecting sentiment polarity of customers' opinions or	Machine Learning Deep Learning Lexicon-based methods Hybrid Methods	(Berlanga, et al., 2015) (Zucco, Calabrese, Agapito, Guzzi, & Cannataro, 2020) (Rosenthal, Farra, & Nakov, 2017) (Chandra Pandey, Singh Rajpoot, & Saraswat, 2017) (Ehsan Basiri, Nemati, Abdar, Cambria, & Acharya, 2021)
Topic Modelling	Identifying the topics mentioned by custom	LDA (Latent Dirichlet Allocation) LSA (Latent Semantic Analysis) Document Embeddings	(Xie, Zhu, Jiang, Lim, & Wang, 2016) (Nolasco & Oliveira, 2016) (Dahal, Kumar, & Li, 2019) (Özyurt & Akcayol, 2021) (Vayansky & Kumar, 2020)
Author Profiling	Age, gender, variety of language, ethnicity. Online reputation management. Individuals vs organizations. Identifying Influencers, promoters, bots. Political affiliation. Hate speech spreaders	Machine Learning Deep Learning	(Nebot, Rangel, Berlanga, & Rosso, 2018) (Wood-Doughty, Mahajan, & Dredze, 2018) (Daelemans, et al., 2019) (Potthast, Rosso, Stamatatos, & Stein, 2019) (Rangel, Sarracén, Chulvi, Fersini, & Rosso, 2021) (Bevendorff, 2023)
Predictive Modelling (Time series)	Retweet prediction, user behaviour prediction, predict stock price, personality-based product recommender, determine cognitive patterns that elicit consumer engagement	Machine Learning Deep Learning	(Schivinski, 2021) (Nasiri, et al., 2017) (Zhang, Shi, Wang, & Fang, 2018) (Ma, Hu, Zhang, Huang, & Jiang, 2019) (Buettner, 2017)
Network-based Analysis	Identifying customers' interaction network or discovering influential users.	Social network analysis. Network diffusion models.	(Peng, et al., 2018) (Peng, Yu, & Mueller, 2018)

Table 1. Main analytical tasks and methods investigated in the academic literature for social analytics and business intelligence

The SBI research area aims at integrating the flow of social data into the BI pipeline to enhance the analytical potential of managers and the decision-making process (Gallinucci, Golfarelli, & Rizzi, 2015) (Francia, Gallinucci, & Golfarelli, 2019). In the business world, SBI-oriented processes are generally implemented through so-called social media monitoring tools, which monitor traditional metrics such as engagement, reach and influence. Popular social media tools rely on fixed sets of dashboards that analyse data from specific perspectives, such as topic usage, topic correlation, and brand reputation. They also rely on ad-hoc KPIs, including topic count and sentiment. However, these approaches do not offer flexible analytical solutions that allow for linking to the organization's internal KPIs (Francia, Gallinucci, Golfarelli, & Rizzi, Social Business Intelligence in Action, 2016). In the context of B2B companies, most studies are conducted using online surveys and content analysis. However, there is still a lack of in-depth research on generalized methodologies focused on SBI in the academic realm (Dwivedi, Ismagilova, Rana, & Raman, 2023).

In (Berlanga, et al., 2015), the authors use opinion mining and sentiment analysis techniques to extract sentiment data from social media for BI applications. They propose a semantic data infrastructure that aims to provide resources for a streamlined integration of internal organizational and social data. The infrastructure follows the principles of the LOD cloud initiative. It provides mechanisms for extracting, linking and publishing social data in the form of RDF triples modelled as multidimensional stars. A semantic meta-structure of multidimensional analysis patterns (user facts, social facts and dimensions) is proposed to model social data. A functional architecture is proposed whose lower layer is responsible for executing the ETLink (Extract, Transform and Link semantically) processes to populate the datasets. The data services layer hosts the services responsible for producing the data required by the business analysis tools. One problem with this approach is that physically storing all the data extracted from social sources as LOD can be costly in terms of storage resources and query time. Therefore, it is necessary to define solutions for streaming data processing and storing only summaries built on the fly.

Another interesting approach was presented in (Francia, Golfarelli, & Rizzi, 2014). The authors propose an architecture and an iterative methodology that provides the basis for designing and maintaining a SBI application. As sources of analysis, they propose the use of social networks, blogs and digital newspapers. The researchers proposed the utilization of several indicators to monitor the various stages of the project. For instance, one of the indicators focuses on analysing the coverage of the ontology, specifically measuring the percentage of clips that include at least one topic from the ontology. Another indicator, known as content relevance analysis, evaluates the effectiveness of the crawling process by assessing the percentage of clips that are relevant to the specified topic. Additionally, the researchers suggest indicators to monitor the improvement of test results at each stage of the methodology following subsequent iterations. Analytical data processing is deferred, combining multidimensional data repositories (Data Warehouse and OLAP) to store textual data and ontologies to represent domain concepts, so it hardly respond to the frequent changes the analysts require. It is worth noting that this research presents a compelling approach to the utilization of specific quality indicators for controlling process improvement, analysing the post collection based on topic coverage, and determining the total number of relevant posts. However, it is important to notice that these indicators are confined to quantitative metrics that do not evaluate the degree or proportion of genuine relevance to a specific topic. Furthermore, an essential perspective for analysis, namely the user profile, who serves as a crucial indicator of content credibility, is overlooked.

In (Abu Salih, Wongthongtham, Beheshti, & Zajabbari, 2019) , the authors propose a conceptual framework for SBI based on the life cycle of big data analysis. They identify three key requirements and major challenges that have not been widely addressed in the existing literature. Firstly, they emphasize the necessity for new systems that can handle the massive and distributed processing of big social data. Secondly, they highlight the importance of implementing mechanisms to assess the reliability of the data. Lastly, they discuss the need for solutions to enrich textual data using semantic representations, including vocabularies, taxonomic relationships, and ontologies for domain knowledge inference.

In (Choi, Yoon, Chung, Coh, & Lee, 2020), the authors conducted a systematic review and examined the analytical methods used in a representative sample of social media BI research. They summarized the key findings in a conceptual framework consisting of four steps, which align with the majority of the reviewed studies: data collection, pre-processing, analysis employing various algorithms, and validation and interpretation of the results. First, data collection refers to the collection of text, image or video data through an open API or web crawling, and the configuration of a database for later analysis. Next, data pre-processing or cleansing is performed to obtain a qualified data set. Specifically, data pre-processing of texts consists of a keyword extraction process using Natural Language Processing. Data analysis involve a variety of data processing and knowledge discovery algorithms such as topic modelling, sentiment analysis and machine learning. Lastly, statistical validation and qualitative interpretation are performed. In this research, various areas for improvement are identified. Firstly, social media BI studies should extend their analysis beyond sentiment analysis by developing quantitative approaches to identify specific reasons for negative reviews and discover new features for customer satisfaction and acquisition in real-world business environments. Secondly, the diverse nature of social media data, encompassing different types of information beyond sentiment scores, presents valuable insights related to product opportunities, emerging technologies, and customer-driven functionalities.

In (Gioti, Ponis, & Panayiotou, 2018) an extensive study of the SBI scientific field is made. They identify three basic pillars of research orientation and categorise them as follows: Business Descriptive (Definitions, Methods, Models & Frameworks), Technical Descriptive (Algorithms, Techniques & Tools) and Case Descriptive (Empirical Evidence / Industry Focus & Social Media Focus). The main shortcomings identified in the first research pillar relate to data security and privacy, and data and process governance. The main shortcomings identified in the second pillar are related to the increased demand for new AI algorithms to automate the extraction of user-generated content, user profiling and the underlying targeted marketing and personalised recommendation systems, as well as a shortcoming in data visualisation systems. Finally, among the main limitations related to the third research pillar, authors mention that the empirical evidence provided in the study is isolated from the general scenario of the company and that the link with result metrics is scarce.

In general, all of these works tackle the primary challenges in the field of SBI, providing partial solutions to various aspects of an SBI project. However, there is a lack of a holistic perspective on the entire problem, and certain crucial aspects like author profiling and data quality management received little attention. In addition, the extensive literature consulted did not address the development of indicators based on social media metadata, metrics, and texts, which is crucial for understanding social behaviour in these platforms. It is necessary to develop methodologies that enable effective social analysis through robust indicators, while also facilitating the connection of these indicators with strategic indicators from other sources to explain and predict social trends or patterns. The analysis of social

metrics is fundamental for measuring the user's reaction to certain actions or events, which in turn enables the analysis of post and user relevance. It is also necessary to automate mechanisms for analysing data quality based on the credibility of user profiles and the analysis context.

1.4 Research context

The increased use of social media has revolutionised the way people interact, shop and engage with brands. In this context, the development of SBI software is an opportunity to improve business decision-making and gain a competitive advantage. However, in the context of social media, data analysis requires a specific and different approach to traditional BI methods due to the unstructured, real-time and complex nature of the data generated on these platforms. Research in this field has focused on the development of data analysis tools and data mining algorithms to improve the understanding of user trends and preferences on social networks. However, there is still a gap in integrating these tools into a comprehensive and easy-to-use solution.

In relation to this new research area, it is worth mentioning that the TKBG (Temporal Knowledge Bases Group) research group at Universitat Jaume I, led by the supervisor of this research, has made several impactful contributions. Firstly, the group has continuously participated in several competitive R&D projects funded by the Ministry of Economy and Commerce and the Generalitat Valenciana. Specifically, the project "Desarrollo de una infraestructura abierta de datos enlazados para la inteligencia de negocio social orientada a PYME" (TIN2014-55335-R) and the project "CogMining: Desarrollo de Técnicas Cognitivas, de Integración Semántica de la Información y de Minería de Datos para Sistemas de Apoyo a la Decisión" (GV/2013/135) are directly relevant to the research line proposed in this thesis. Two other important scientific contributions are the postdoctoral project presented by Victoria Nebot in (Nebot & Berlanga, 2016) (Nebot, Rangel, Berlanga, & Rosso, 2018) and the doctoral thesis of Lisette García presented in (García Moya, 2016). Both research studies have made significant contributions to the development of the SLOD-BI (Social Linked Open Data for Business Intelligence) infrastructure, which serves as an open data framework for facilitating SBI (Berlanga, et al., 2015). These projects serve as a prelude to this thesis proposal, as they have established the essential groundwork for the development of new and closely related research lines.

SLOD-BI proposes a LOD infrastructure aimed at capturing and publishing facts extracted from social networks that are relevant to the strategic objectives of SME companies. In my Master's thesis (Lanza Cruz I. , Definición y análisis de indicadores estratégicos para redes sociales : un caso de estudio en el sector automovilístico, 2016), I previously presented a proposal for representing social indicators within the SLOD-BI infrastructure using state-of-the-art ontologies that provide a logical framework for knowledge representation. Following the principle of the infrastructure, the proposal designs new components that extend the BI patterns developed in SLOD-BI. New functionalities have been added, including the calculation of social indicators based on the social measures captured by the platform, as well as the definition of the necessary dimensions for analysis.

Figure 2 shows the main components of the SLOD-BI infrastructure. The inner ring of the figure comprises the components (include vocabularies and datasets) of the infrastructure, while the outer ring includes two sets: one consisting of external

vocabularies expressed as linked open vocabularies (In the figure, they are represented in boxes with dashed lines), and the other comprising external datasets that can be linked to the infrastructure, such as DBpedia.org and BabelNet.org. Each SLOD-BI component consists of one or more datasets representing one of the perspectives considered relevant for BI analysis of opinion data. The links represented by the thicker lines must be semantically coherent, as they are often used to perform analysis tasks. Semantic coherence in the context of the semantic web refers to the clear and controlled definition of ontologies and vocabularies to establish a common framework of meanings. Consistency in the use of terms must be maintained throughout the dataset, and the relationships between terms should be logical and consistent in the context of information. Naturally, established semantic standards must be followed for the semantic representation of information, such as RDF and OWL. Conversely, the links denoted by thinner lines suggest potential connections between entities within the infrastructure and external datasets. Finally, the direction of the arrows represents potential dependencies from one component to another.

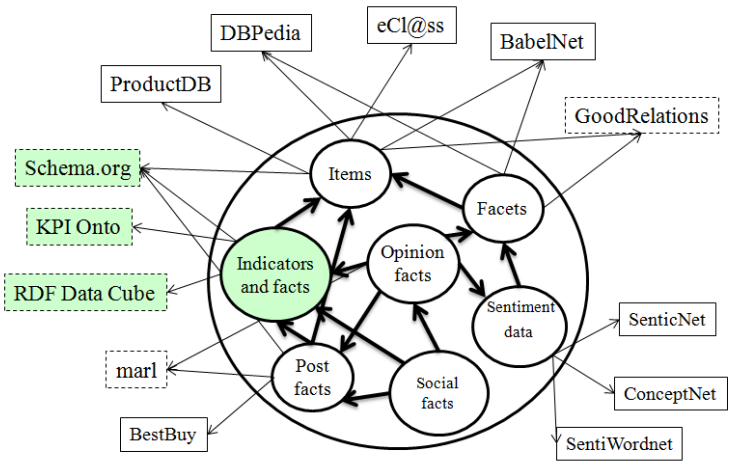


Figure 2 – SLOD-BI Data Infrastructure **(Lanza Cruz I. , Definición y análisis de indicadores estratégicos para redes sociales : un caso de estudio en el sector automovilístico, 2016)**

The main limitation of this infrastructure was its static nature, as the data was generated and published as static RDF datasets. However, the social facts generated in social networks are highly dynamic and often need to be analysed in near real time (right time). For these reasons, this work faces new challenges, such as the dynamic generation of knowledge, maintaining its coherence, and ensuring the quality and relevance of the results as much as possible.

1.5 Limitations and Open Issues

In today's dynamic business landscape, companies recognize the well-established benefits that social network analytics bring to business development. As a result, there is a growing advocacy for the standardization of methodologies to enhance the effectiveness of social network analysis and maximize its impact on strategic decision-making. However, existing proprietary software tools for social analytics primarily focus on analysing commonly used social metrics like engagement, reach, influence, and sentiment. They often lack the capability to establish robust connections between these metrics and KPIs that evaluate the achievement of business goals. In addition, due to their proprietary nature, AI tools for social analytics are “black boxes” that force users to accept results on blind faith, a source of concern in industry and academia (L. Hayes, et al., 2020). In the academic field, most proposals for social data analysis focus on solving problems in closed contexts and do not solve the problem of data dynamicity, variety and speed. For example, many approaches create ad-hoc processes to measure some type of indicator limited to a specific topic or dimension, mainly topological (Wang, Jiao, Abrahams, Fan, & Zhang, 2013.), product (Yan, Xing, Zhang, & Ma, 2015) or sentiment (Dai, Han, Dai, & Xu, 2015). For these reasons, future research should evolve towards a systematic methodology beyond a single static analysis (Choi, Yoon, Chung, Coh, & Lee, 2020).

In the scientific literature, there are few references based on the development of integral methodologies and frameworks for SBI. The reviewed solutions in Section 1.2 show the following limitations and open issues, which will be addressed in this PhD research:

- **[Social indicators using semantics models]** The review has revealed that there are currently no comprehensive solutions available for modelling social indicators with the purpose of evaluating the social response to marketing efforts and organizational strategies in social networks. It is imperative to establish mechanisms for defining social indicators using semantic models to represent linked data (LD) and enable multidimensional exploration. Furthermore, the semantic infrastructure must serve as a foundation for several analytical approaches, including author profiling, quality analysis, predictive analytics, and more. This will assist organizations in anticipating changing market or social conditions, as well as identifying potential markets and customers. Finally, it is crucial for the solution to support dynamic analysis.

- **[Author Profiling]** The role of the user is a fundamental analytical pattern in BI projects. However, in current SBI projects, this perspective has received a partial or null approach. It is crucial to have solutions that enable author profiling in social networks, using dynamic analysis dimensions, to understand the user's needs, which are at the core of the analysis. It is also necessary to develop author profile qualitative indicators in order to assess the reliability of the collected posts.

- [**Quality management]** When it comes to data quality management in SBI projects, previous frameworks have not sufficiently emphasized the significance of data quality during the data preparation phases. The review highlights the absence of comprehensive, general-purpose conceptual models for evaluating quality in social media data, underscoring the necessity for systematic methodologies to identify and

apply appropriate quality metrics for filtering out irrelevant or noisy content. It is imperative to implement effective methods for noise elimination and to identify data and value metrics that align with analysis objectives.

- **[Data stream processing]** Information in big data environments, such as social networks, evolves rapidly, requiring real-time streaming data analysis and on-the-fly summarization of social indicators to avoid costly storage operations. Currently, there is a lack of solutions that meet these requirements. It is crucial to develop general mechanisms for effectively harnessing big data, overcoming its unstructured and fragmented nature.

1.6 Hypothesis

Building upon the context and problem scenario described in the previous section, this thesis proposes the development of novel dynamic indicators and dimensions derived from social media data to assist in decision-making within SBI projects. For this purpose, we develop a methodology to formally define social indicators, in turn, the implementation of an infrastructure that allows their management and easy integration with the SBI systems in an efficient and scalable way. We state the following hypotheses to limit the scope of this thesis.

Hypothesis 1

The implementation of a framework based on semantic models for representing social indicators, in conjunction with the development of effective methods for processing real-time social media data, has a positive impact on the more precise identification and tracking of social patterns, behaviours, and trends. The linked data infrastructure provides a more structured organization of social metrics and metadata, facilitating semantic enhancement of captured information. Furthermore, the use of linked data enables the discovery of new patterns and relationships in the data, thereby contributing to more efficient modelling of social indicators.

Hypothesis 2

The implementation of an author profiling method based on multidimensional business perspectives, derived from authors' biographies on social networks, is positively associated with the accuracy and reliability in identifying authors' business roles. The qualitative information resulting from this profiling will facilitate user segmentation tailored to analysis needs. Furthermore, it is hypothesized that author profiling can serve as a novel quality indicator enabling the evaluation of content reliability.

Hypothesis 3

It is feasible to develop quality indicators semi-automatically to identify relevant content on social media that meets dynamic analysis criteria. Evaluating data quality in a social media collection involves the analysis of two main types of indicators: author profile indicators and context-specific quality metrics. The effective combination of these quality

indicators has the potential to significantly enhance data collections, enabling more effective data analysis and more precise decision-making.

1.7 Goals

The aim pursued by this thesis is the development of a comprehensive methodological framework for defining and monitoring social indicators in SBI projects. Derived from this overarching objective, the following goals are proposed:

Goal 1. To define a methodological framework for the design, specification and monitoring of social indicators based on social info-metrics and dynamic domain topics. The associated methods must support the processing of data streams to feed the defined indicators. The data should be modelled using the multidimensional model to facilitate analysis, visualization, and understanding from different perspectives. To streamline data processing and facilitate data integration, this model should be implemented using semantic technologies. This objective is aligned with hypothesis 1, and it has been successfully achieved in the following publications:

- Lanza Cruz, Indira and Berlanga Llavori, Rafael. (2018). Defining Dynamic Indicators for Social Network Analysis: A Case Study in the Automotive Domain using Twitter. *In Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR*, ISBN 978-989-758-330-8; ISSN 2184-3228, pages 221-228. DOI: 10.5220/0006932902210228.

- Lanza Cruz, Indira; Berlanga, Rafael and Aramburu, María José. (2018). Modeling Analytical Streams for Social Business Intelligence. *Informatics*, 5, 33. doi.org/10.3390/informatics5030033

Goal 2. To develop a method to specify user indicators, particularly qualitative indicators based on the user's business role, by leveraging automatic classification of social network's user profiles. This method should capture the multidimensional nature and dynamic perspectives that a SBI project requires. This objective is aligned with hypothesis 2, and it has been successfully achieved in the following publication:

- Lanza Cruz, Indira; Berlanga, Rafael and Aramburu, María José. (2023). Multidimensional Author Profiling for Social Business Intelligence. *Information Systems Frontiers*. Springer. doi.org/10.1007/s10796-023-10370-0.

Goal 3. To define a method for the automatic generation of quality indicators to support the construction of high-quality datasets adapted to the context of analysis in an SBI project. This objective is aligned with hypothesis 3, and it has been successfully achieved in the following publications:

- Berlanga, R.; Lanza Cruz, Indira and Aramburu, María José. (2019). Quality Indicators for Social Business Intelligence. *In 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Granada, Spain, 2019, pp. 229-236, doi: 10.1109/SNAMS.2019.8931862.

- Aramburu, María José; Berlanga, Rafael and Lanza Cruz, Indira. (2023). A Data Quality Multidimensional Model for Social Media Analysis. *Business & Information Systems Engineering*, DOI: 10.1007/s12599-023-00840-9 (In Press).

Goal 4. To define an evaluation framework for the proposal that demonstrates the quality of the data used to feed the indicators, the effectiveness of the indicators in their strategic purpose, and the reliability of the proposed predictive methods. This objective has been successfully achieved in the following publication:

- Aramburu, María José; Berlanga, Rafael; Lanza Cruz, Indira. (2020). Social Media Multidimensional Analysis for Intelligent Health Surveillance. *International Journal of Environmental Research and Public Health*, 17, 2289. doi.org/10.3390/ijerph17072289.

1.8 Methodology

In this PhD. thesis, the fundamental stages of scientific research development have been diligently followed, encompassing the definition of research questions, hypotheses, and objectives, an extensive review of the scientific-technical background in the research area, the proposal of novel methods that effectively address the primary objectives, and ultimately, the validation of the proposed approach through experiments and proof of concept.

The main purpose of this research is to develop an effective set of social indicators that enable the evaluation of an organization's impact on social networks through its actions. To achieve this objective, a methodological approach based on several steps was followed. Firstly, a comprehensive review of the existing literature on performance indicators in the company was conducted. Special attention was given to the study of formalized methodologies and models, such as the BIM model (Business Intelligence Model) (Horkoff, Barone, & Jiang, 2014), as well as frameworks and ontologies used for their definition and monitoring. In addition, methods for selecting social media metrics aimed at measuring customer experience and evaluating marketing strategies on social networks were studied. The most commonly used analytical tasks during social network analysis were also examined, with a particular focus on techniques employed for author profiling and existing methods of data quality management. All of this was done with the aim of adopting a comprehensive perspective during the specification and development of social indicators. This review provided a solid theoretical foundation and allowed for the identification of best practices and approaches used in measuring actions and performance on social networks.

To address the general goal pursued by this thesis, we followed generic methodologies for the design, development, and deployment of SBI projects. These methodologies

included general phases such as business understanding and requirements capture, data extraction and transformation, data model design, and data exploitation through the creation of reports and dashboards. Appendix “A” presents a diagram of the followed methodology, wherein each stage connects the various methods proposed in this thesis.

To address the different solutions, multidimensional and dynamic analysis approaches were applied in a general manner. The aim was to provide generic solutions that can be adapted to the required dynamicity of the analysis needs of SBI projects. Specifically, methodologies for ontology development and knowledge graphs were employed in the specification of data models. The use of semantic models through LD and multidimensional models in representing indicators, offers several significant advantages. These include the flexibility to make quick changes and adaptations compared to traditional models, making them highly useful in business environments that require agile and adaptable responses to changing requirements. Additionally, they enable the addition of context and meaning to indicators by linking them to concepts and relationships defined in the ontologies. They also facilitate information reuse, ease of querying, and analysis. To implement and process these data models, formalized models and techniques from other studies, e.g. (Nebot & Berlanga, 2016), were followed.

For author profiling, we propose combining an ontological model of user roles with text classifiers applied to user descriptions. The ontological model aims to establish a hierarchical and multidimensional perspective of user roles, along with a set of constraints that ensure compatibility between roles. This model is integrated within the Dynamic SLOD-BI infrastructure. Text classifiers have been implemented to assign multidimensional business role labels to each user profile. In our proposal, we employ an unsupervised approach to generate labelled datasets using bigrams and semantic embeddings derived from the ontology of business roles.

For the development of quality indicators, we give special importance to the contents coherence and the data provenance as fundamental quality criteria. This required conducting a critical analysis of existing quality management methodologies for processing social media data. To refine and validate the quality indicators, information retrieval methods and a reference dataset were employed.

In parallel, several use cases related to important economic sectors were developed to validate the proposed solutions. For this purpose, we evaluated our approaches using long-term streams of Twitter data specific to each domain.

This comprehensive methodology allowed us the development of suitable and customized social indicators to meet the analysis needs. These indicators will serve as a strong basis for measuring the quality of social data and assessing its social impact, thereby supporting informed strategic decision-making.

1.9 Ethical, Legal, and Bias Handling Issues

1.9.1 Ethical and Legal questions

During the development of this thesis, the importance of complying with Twitter's legal terms and conditions in the context of the research was emphasized. This is essential for ethical and legal reasons. Privacy and user rights were respected by anonymizing data and reformulating (using text normalization techniques) the texts of tweets and user descriptions. Compliance with Twitter's policies was ensured, and restrictions in research were addressed, avoiding the inference of sensitive data. As per Twitter's policies, the analyses were conducted on aggregated content that does not store any personal data. We had to consider legal issues, specifically, sharing Twitter datasets is prohibited by Twitter's API Terms of Service. There are some exceptions; for academic research purposes, it is allowed to share only the identification numbers (IDs) of the tweet or the user with the aim of reproducing the experiments. However, in this case, researchers may choose not to publicly disclose the tweet IDs due to the potential risk of harm. A solution could be for academic institutions to formally request the exchange of these data for research purposes (X Corp., 2023).

Emphasis was placed on equity and non-discrimination, with algorithms adjusted to avoid unwanted biases. Specifically, no preference has been given in any data condition, for example socio-demographic data is not used. Transparency and disclosure were fundamental, describing how the data and analysis techniques were used. Additionally, copyright was respected, and appropriate attributions were provided.

1.9.2 Bias Handling Issues

In addition to the ethical considerations we faced, there were also challenges related to data acquisition. It is essential to understand the limitations of Twitter data and their impact on fair analysis and the configuration of machine learning models.

In this thesis, we address the issue of bias by starting with a thorough analysis of the patterns in the data that would feed our machine learning models. During the design of the author profiling model, we extensively explore the datasets resulting from different iterations of automated labelling processes to achieve a balance in the distribution of data features. To do this, we analyse class distribution and adjust the number of key bigrams when the minority class has insufficient samples. The resulting models were evaluated with both balanced and unbalanced class distributions, yielding similar scores. Additionally, we conduct an analysis of feature imbalance using data dispersion analysis techniques and exploration of the vector space, applying feature selection techniques such as TSNE and PCA. We also employ techniques for undersampling similar key bigrams (or keywords) to create more diverse sets within the same class. The proposed technique for conflict resolution during profile labelling allowed for adjusting the proportion of examples from different classes containing the same key bigram in the training dataset to balance the classes.

It's worth noting that we have utilized embeddings from language models like FastText and RoBERTa (Lanza Cruz, Berlanga, & Aramburu, Multidimensional Author Profiling for Social Business Intelligence, 2023), which provide significant benefits for data generalization and bias mitigation. These models are trained on diverse texts, generating

pretrained embeddings, and tend to capture general linguistic information, reducing the propagation of biases inherent in training data.

Furthermore, to detect biases in data generated by Twitter, it is crucial to conduct a long-term analysis, as biases can be subtle and not evident in the short term. Continuous data collection over time allows for the identification of stable patterns and differentiation between temporal biases and persistent bias patterns. This facilitates the ongoing improvement of models and bias mitigation strategies.

Additionally, we acknowledge the potential for missing data during data acquisition and how this can lead to a biased sample. Language choice can also influence data representativeness. In our case, the Twitter data collected is mainly in English.

It is important to note that Twitter users do not necessarily represent the general population or even all internet users, and Twitter data may not be representative of Twitter users themselves due to the specific nature of topics discussed in tweets.

1.10 Outline of the PhD Thesis Report

The research carried out in this thesis has been motivated by the problems and open lines of research discussed in the previous sections. The thesis is presented by means of a compendium of scientific publications that form part of the research results and are presented as chapters. The manuscript is divided into 7 chapters, including this one, where an introductory chapter is presented.

The chapters 2 to 6 present the papers selected to be part of this thesis. The next two chapters address the Goal 1. Specifically, Chapter 2 presents the paper "Defining Dynamic Indicators for Social Network Analysis: A Case Study in the Automotive Domain using Twitter" (Lanza Cruz & Berlanga Llavori, 2018). This chapter presents a framework based on LD technologies to perform analysis tasks in social networks based on dynamically defined social indicators. Chapter 3 presents the paper "Modelling Analytical Streams for Social Business Intelligence" (Lanza Cruz, et al., 2018). This work proposes a multidimensional formalism to represent and evaluate social indicators directly from social fact streams. Chapter 4 responds to Goal 2 and presents the work "Multidimensional Author Profiling for Social Business Intelligence" (Lanza Cruz, et al., 2023). The research presents a novel author profiling method aimed at classifying social network users into the multidimensional perspectives for SBI applications.

In the context of this thesis, we address the design of a semantic and dynamic infrastructure for managing social indicators. However, it is crucial to highlight that an indicator can only be considered valid when it is fuelled by high-quality data. Therefore, the development of this infrastructure requires the implementation of efficient mechanisms to assess the quality of the data it manages. To meet these requirements, we have defined Goal 3, which is developed on in the detailed research presented in Chapter 5. This chapter presents two articles that constitute a sequence of enhancements to the research developed. As a result, it proposes novel methods for building high quality datasets for SBI projects based on the semi-automatic definition of quality indicators. The related papers are "Quality Indicators for Social Business Intelligence" (Berlanga, et al., 2019) and "A Data Quality Multidimensional Model for Social Media Analysis" (Aramburu, Berlanga, & Lanza-Cruz, A Data Quality Multidimensional Model for Social Media Analysis, 2023).

In order to address the issues related to Goal 4, Chapter 6 presents the paper "Social Media Multidimensional Analysis for Intelligent Health Surveillance" (Aramburu,

Berlanga, & Lanza, 2020). This last research aims to integrate the results obtained in the previous research works in the same semantic infrastructure and to validate the proposed methodologies by developing a practical use case related to the domain of surveillance in public health from social networks.

The last chapter contains the general discussion and conclusions sections. This chapter summarises the main contributions and results that validate the hypotheses. In a subsequent section, the scientific publications developed during this thesis and derived research projects are related. Finally, the chapter outlines and describes interesting lines for further research.

References

Abu Salih, B., Wongthongtham, P., Beheshti, S., & Zajabbari, B. (2019). Towards a Methodology for Social Business Intelligence in the Era of Big Social Data Incorporating Trust and Semantic Analysis. *Proceedings of the International Conference on Data Engineering, 520*. doi:https://doi.org/10.1007/978-981-13-1799-6_54

Allemang, D., & Hendler, J. (2011). *Semantic Web for the Working Ontologist : Effective Modeling in RDFS and OWL*. Elsevier Science & Technology.

Aramburu, M., Berlanga, R., & Lanza Cruz, I. (2021). Quality Management in Social Business Intelligence Projects. *International Conference on Enterprise Information Systems, 1*, pp. 320-327. doi:10.5220/0010495703200327

Aramburu, M., Berlanga, R., & Lanza, I. (2020). Social Media Multidimensional Analysis for Intelligent Health Surveillance. *Int. J. Environ. Res. Public Health, 17*. doi:https://doi.org/10.3390/ijerph17072289

Aramburu, M., Berlanga, R., & Lanza-Cruz, I. (2023). A Data Quality Multidimensional Model for Social Media Analysis. *Business & Information Systems Engineering*. doi: https://doi.org/10.1007/s12599-023-00840-9

Berlanga, R., García-Moya, L., Nebot, V., Aramburu, M., Sanz, I., & Llidó, D. (2015). SLOD-BI: An Open Data Infrastructure for Enabling Social Business Intelligence. *International Journal of Data Warehousing and Mining (IJDWM), 11*(4), 1-28. doi:10.4018/ijdwm.2015100101

Berlanga, R., Lanza Cruz, I., & Aramburu, M. J. (2019). Quality Indicators for Social Business Intelligence. *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 229-236. doi:http://dx.doi.org/10.1109/SNAMS.2019.8931862

Bevendorff, J. e. (2023). Overview of PAN 2023: Authorship Verification, Multi-author Writing Style Analysis, Profiling Cryptocurrency Influencers, and Trigger Detection. *In: Kamps, J., et al. Advances in Information Retrieval. ECIR 2023. Lecture Notes in Computer Science, 13982*. doi:https://doi.org/10.1007/978-3-031-28241-6_60

Buettner, R. (2017). Predicting user behavior in electronic markets based on personality-mining in large online social networks. *Electron Markets, 27*, 247–265. doi:https://doi.org/10.1007/s12525-016-0228-z

Chandra Pandey, A., Singh Rajpoot, D., & Saraswat, M. (2017). Twitter sentiment analysis using hybrid cuckoo search method. *Information Processing & Management, 53*(4), 764-779. doi:https://doi.org/10.1016/j.ipm.2017.02.004

Choi, J., Yoon, J., Chung, J., Coh, B.-Y., & Lee, J.-M. (2020). Social media analytics and business intelligence research: A systematic review. *Information Processing & Management, 57*(6). doi:https://doi.org/10.1016/j.ipm.2020.102279

Coyne, S., Rogers, A., Zurcher, J., Stockdale, L., & Booth, M. (2020). Does time spent using social media impact mental health?: An eight year longitudinal study. *Computers in Human Behavior, 104*. doi:https://doi.org/10.1016/j.chb.2019.106160.

Daelemans, W., Kestemont, M., Manjavacas, E., Potthast, M., Rangel, F., Rosso, P., . . . Zangerle, E. (2019). Overview of PAN 2019: bots and gender profiling, celebrity profiling, cross-domain authorship attribution and style change detection. *F. Crestani, et al. (Eds.), Lecture notes in Computer Science, vol11696,*

experimental IR meets multilinguality, multimodality, and Interaction. doi: https://doi.org/10.1007/978-3-030-28577-7_30

Dahal, B., Kumar, S., & Li, Z. (2019). Topic modeling and sentiment analysis of global climate change tweets. *Soc. Netw. Anal. Min., 9*. doi:https://doi.org/10.1007/s13278-019-0568-8

Dai, W., Han, D., Dai, Y., & Xu, D. (2015). Emotion recognition and affective computing on vocal social media. *Information & Management, 52*(7), 777-788. doi:https://doi.org/10.1016/j.im.2015.02.003

Diamantini, C., Potena, D., & Storti, E. (2016). SemPI: A semantic framework for the collaborative construction and maintenance of a shared dictionary of performance indicators. *Future Generation Comp. Syst., 54*, 352-365. doi:https://doi.org/10.1016/j.future.2015.04.011

Dwivedi, Y. K., Ismagilova, E., Rana, N. P., & Raman, R. (2023). Social Media Adoption, Usage And Impact In Business-To-Business (B2B) Context: A State-Of-The-Art Literature Review. *Information Systems Frontiers, 25*. doi:https://doi.org/10.1007/s10796-021-10106-y

Ehsan Basiri, M., Nemati, S., Abdar, M., Cambria, E., & Acharya, U. R. (2021). ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis. *Future Generation Computer Systems, 115*. doi:https://doi.org/10.1016/j.future.2020.08.005.

Francia, M., Gallinucci, E., & Golfarelli, M. (2019). Social BI to understand the debate on vaccines on the Web and social media: unraveling the anti-, free, and pro-vax communities. *Italy. Soc. Netw. Anal. Min., 9*(46). doi:https://doi.org/10.1007/s13278-019-0590-x

Francia, M., Gallinucci, E., Golfarelli, M., & Rizzi, S. (2016). Social Business Intelligence in Action. (Springer, Ed.) *Advanced Information Systems Engineering. CAiSE 2016. Lecture Notes in Computer Science, 9694*. doi:https://doi.org/10.1007/978-3-319-39696-5_3

Francia, M., Golfarelli, M., & Rizzi, S. (2014). A Methodology for Social BI. *Proceedings of the 18th International Database Engineering & Applications Symposium. IDEAS '14*, 207–216. doi:https://doi.org/10.1145/2628194.2628250

Gallinucci, E., Golfarelli, M., & Rizzi, S. (2015). Advanced topic modeling for social business intelligence. *Information Systems, 53*, 87–106. doi:https://doi.org/10.1016/j.is.2015.04.005

García Moya, L. (2016). *Modeling and analyzing opinions from customer reviews.* (U. J. I, Ed.) doi:http://dx.doi.org/10.6035/40009.2016.247757

Gioti, H., Ponis, S. T., & Panayiotou, N. (2018). Social business intelligence: Review and research directions. *Journal of Intelligence Studies in Business, 8*(2). doi:https://doi.org/10.37380/jisib.v8i2.320

Heath, T., & Bizer, C. (2022). *Linked data: Evolving the web into a global data space.* Springer Nature.

Horkoff, J., Barone, D., & Jiang, L. e. (2014). Strategic business modeling: representation and reasoning. *Softw Syst Model, 13*, 1015–1041. doi:https://doi.org/10.1007/s10270-012-0290-8

Jiménez-Ruiz, E., Hassanzadeh, O., Efthymiou, V., Chen, J., & Srinivas, K. (2020). SemTab 2019: Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems. *Harth, A., et al. The Semantic Web. ESWC 2020. Lecture Notes in Computer Science, 12123*. doi:https://doi.org/10.1007/978-3-030-49461-2_30

Knoke, D., & Yang, S. (2019). *Social network analysis.* SAGE publications.

L. Hayes, J., Britt, B. C., Evans, W., Rush, S. W., Towery, N. A., & C., A. (2020). Can Social Media Listening Platforms' Artificial Intelligence Be Trusted? Examining the Accuracy of Crimson Hexagon's (Now Brandwatch Consumer Research's) AI-Driven Analyses. *Journal of Advertising, 50*(1), 81-91. doi:10.1080/00913367.2020.1809576

Lanza Cruz, I. (2016). *Definición y análisis de indicadores estratégicos para redes sociales : un caso de estudio en el sector automovilístico.* Trabajo de Fin de Máster, Universitat Jaume I, Castellón de La Plana. Retrieved from http://hdl.handle.net/10234/165315

Lanza Cruz, I. L., & Berlanga Llavori, R. (2018). Defining Dynamic Indicators for Social Network Analysis: A Case Study in the Automotive Domain using Twiter. (SciTePress, Ed.) *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management. KDIR, 1*, 221-228. doi:10.5220/0006932902210228

Lanza Cruz, I., Aramburu, M. J., & Berlanga, R. (2016). Metodología de inteligencia de negocio para análisis social en la infraestructura de datos enlazados SLOD-BI. *Ciência da Informação. Informação estratégica , 45*(3). doi: https://doi.org/10.18225/ci.inf.v45i3.4058

Lanza Cruz, I., Berlanga, R., & Aramburu, M. J. (2018). Modeling Analytical Streams for Social Business Intelligence. *Informatics, 5*(3). doi:https://doi.org/10.3390/informatics5030033

Lanza Cruz, I., Berlanga, R., & Aramburu, M. J. (2023). Multidimensional Author Profiling for Social Business Intelligence. *Inf Syst Front*. doi:https://doi.org/10.1007/s10796-023-10370-0

Liu, X., Tang, K., Hancock, J., Han, J., Song, M., Xu, R., & Pokorny, B. (2013). A Text Cube Approach to Human, Social and Cultural Behavior in the Twitter Stream. (Springer, Ed.) *Greenberg, A.M., Kennedy, W.G., Bos, N.D. (eds) Social Computing, Behavioral-Cultural Modeling and Prediction. SBP 2013. Lecture Notes in Computer Science, 7812*. doi:https://doi.org/10.1007/978-3-642-37210-0_35

Ma, R., Hu, X., Zhang, Q., Huang, X., & Jiang, Y. (2019). Hot Topic-Aware Retweet Prediction with Masked Self-attentive Model. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. doi:10.1145/3331184.3331236

Maté, A., Trujillo, J., & Mylopoulos, J. (2012). Conceptualizing and Specifying Key Performance Indicators in Business Strategy Models. *Atzeni, P., Cheung, D., Ram, S. (eds) Conceptual Modeling. ER 2012. Lecture Notes in Computer Science, 7532*. doi:https://doi.org/10.1007/978-3-642-34002-4_22

Maté, A., Trujillo, J., & Mylopoulos, J. (2017). Specification and derivation of key performance indicators for business analytics: A semantic approach. *Data & Knowledge Engineering, 108*, 30-49. doi:https://doi.org/10.1016/j.datak.2016.12.004.

Mena Roa, M. (2023, Abril 25). *Statista*. Retrieved from https://es.statista.com/grafico/18988/tiempo-medio-diario-de-conexion-a-una-red-social/

Nasiri, A., Nalchigar, S., Yu, E., Ahmed, W., Wrembel, R., & Zimányi, E. (2017). From Indicators to Predictive Analytics: A Conceptual Modelling Framework. (Springer, Ed.) *Poels, G., Gailly, F., Serral Asensio, E., Snoeck, M. (eds) The Practice of Enterprise Modeling. PoEM 2017. Lecture Notes in Business Information Processing, 305*. doi:https://doi.org/10.1007/978-3-319-70241-4_12

Nebot, V., & Berlanga, R. (2016). Statistically-driven generation of multidimensional analytical schemas from linked data. *Knowl.-Based Syst., 110*, 15–29. doi:https://doi.org/10.1016/j.knosys.2016.07.010

Nebot, V., Rangel, F., Berlanga, R., & Rosso, P. (2018). Identifying and Classifying Influencers in Twitter only with Textual Information. doi:doi: 10.1007/978-3-319-91947-8_3

Nolasco, D., & Oliveira, J. (2016). Detecting Knowledge Innovation through Automatic Topic Labeling on Scholar Data. *49th Hawaii International Conference on System Sciences (HICSS)*, 358-367. doi:10.1109/HICSS.2016.51

Özyurt, B., & Akcayol, M. A. (2021). A new topic modeling based approach for aspect extraction in aspect based sentiment analysis: SS-LDA. *Expert Syst. Appl., 168*. doi:https://doi.org/10.1016/j.eswa.2020.114231

Parveen Tajudeen, F., Ismawati Jaafar, N., & Ainin, S. (2017). Understanding the impact of social media usage among organizations. *Inf. Manag., 55*, 308-321. doi:10.1016/j.im.2017.08.004

Peng, S., Yu, S., & Mueller, P. (2018). Social networking big data: Opportunities, solutions, and challenges. *Future Generation Computer Systems, 86*, 1456-1458. doi:https://doi.org/10.1016/j.future.2018.05.040.

Peng, S., Zhou, Y., Cao, L., Yu, S., Niu, J., & Jia, W. (2018). Influence analysis in social networks: A survey. *Journal of Network and Computer Applications, 106*, 17-32. doi:https://doi.org/10.1016/j.jnca.2018.01.005

Potthast, M., Rosso, P., Stamatatos, E., & Stein, B. (2019). A decade of shared tasks in digital text forensics at PAN. *L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, & D. Hiemstra, (Eds.), Lecture Notes in Computer Science, 11438*. doi:https://doi.org/10.1007/978-3-030-15719-7_39

Rangel, F., Sarracén, G. L., Chulvi, B., Fersini, E., & Rosso, P. (2021). Profiling hate speech spreaders on Twitter Task. *PAN 2021. CLEF, CEUR-WS.org.*

Roldán-García, M., García-Nieto, J., Maté, A., Trujillo, J., & Aldana-Montes, J. (2021). Ontology-driven approach for KPI meta-modelling, selection and reasoning. *International Journal of Information Management, 102018*. doi:https://doi.org/10.1016/j.ijinfomgt.2019.10.003

Rosenthal, S., Farra, N., & Nakov, P. (2017). SemEval-2017 Task 4: Sentiment Analysis in Twitter. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 502–518. doi:10.18653/v1/S17-2088

Ruhi, U. (2014). Social Media Analytics as a Business Intelligence Practice: Current Landscape & Future Prospects. *Journal of Internet Social Networking & Virtual Communities*. doi:10.5171/2014.920553

Salvatore, C., Biffignandi, S., & Bianchi, A. (2021). Social Media and Twitter Data Quality for New Social Indicators. *Soc Indic Res, 156*, 601–630. doi: https://doi.org/10.1007/s11205-020-02296-w

Schivinski, B. (2021). Eliciting brand-related social media engagement: A conditional inference tree framework. *Journal of Business Research, 130*, 594-602. doi:https://doi.org/10.1016/j.jbusres.2019.08.045

Taulé, M., Rangel, F., Martí, M. A., & Rosso, P. (2018). Overview of the task on multimodal stance detection in tweets on catalan# 1oct referendum. *IberEval@ SEPLN*, 149-166.

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems, 94*. doi:https://doi.org/10.1016/j.is.2020.101582

Wang, G. A., Jiao, J., Abrahams, A. S., Fan, W., & Zhang, Z. (2013.). ExpertRank: A topic-aware expert finding algorithm for online knowledge communities. *Decision Support Systems, 54*(3), 1442-1451. doi:https://doi.org/10.1016/j.dss.2012.12.020

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications* (Vol. 8). (C. U. Press, Ed.)

Wood-Doughty, Z., Mahajan, P., & Dredze, M. (2018). Johns Hopkins or johnny-hopkins: classifying individuals versus organizations on Twitter. *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, 56–61. doi:10.18653/v1/W18-1108

X Corp. (2023). *Twitter - X Developer Platform*. Retrieved 11 13, 2023, from https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases

Xie, W., Zhu, F., Jiang, J., Lim, E. -P., & Wang, K. (2016). TopicSketch: Real-Time Bursty Topic Detection from Twitter. *IEEE Transactions on Knowledge and Data Engineering, 28*(8), 2216-2229. doi:10.1109/TKDE.2016.2556661

Yan, Z., Xing, M., Zhang, D., & Ma, B. (2015). EXPRS: An extended pagerank method for product feature extraction from online consumer reviews. *Information & Management, 52*(7), 850-858. doi:10.1016/j.im.2015.02.002

Yanfang, N., Ying, L., Yang, J., Bao, M., & Sivaparthipan, C. (2021). Organizational business intelligence and decision making using big data analytics. *Information Processing & Management, 58*(6). doi:https://doi.org/10.1016/j.ipm.2021.102725

Zhang, X., Shi, J., Wang, D., & Fang, B. (2018). Exploiting investors social network for stock prediction in China's market. *Journal of Computational Science, 28*, 294-303. doi:https://doi.org/10.1016/j.jocs.2017.10.013

Zucco, C., Calabrese, B., Agapito, G., Guzzi, P., & Cannataro, M. (2020). Sentiment analysis for mining texts and social networks data:Methods and tools. *WIREs Data Mining Knowl Discov*. doi: https://doi.org/10.1002/widm.1333

Chapter 2

Defining Dynamic Indicators for Social Network Analysis: A Case Study in the Automotive Domain using Twitter

Publication

Lanza Cruz, Indira Lázara and Berlanga, Rafael. (2018). Defining Dynamic Indicators for Social Network Analysis: A Case Study in the Automotive Domain using Twitter. In Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KDIR, ISBN 978-989-758-330-8; ISSN 2184-3228, pages 221-228. DOI: 10.5220/0006932902210228.

Defining Dynamic Indicators for Social Network Analysis: A Case Study in the Automotive Domain using Twitter

Indira Lázara Lanza Cruz and Rafael Berlanga Llavori
Llenguatges I Sistemes Informatics, Universitat Jaume I, Castellón, Spain
{lanza, berlanga}@uji.es

Keywords: Social Business Intelligence, Indicators, Data Streaming.

Abstract: In this paper we present a framework based on Linked Open Data Infrastructures to perform analysis tasks in social networks based on dynamically defined indicators. Based on the typical stages of business intelligence models, which starts from the definition of strategic goals to define relevant indicators (Key Performance Indicators), we propose a new scenario where the sources of information are the social networks. The fundamental contribution of this work is to provide a framework for easily specifying and monitoring social indicators based on the measures offered by the APIs of the most important social networks. The main novelty of this method is that all the involved data and information is represented and stored as Linked Data. In this work we demonstrate the benefits of using linked open data, especially for processing and publishing company-specific social metrics and indicators.

1 INTRODUCTION

The main objective of Business Intelligence (BI) is to extract strategic knowledge from the information provided by performance indicators. This knowledge is the basis for facilitating the decision-making process and improving performance in the organization. A performance indicator is used to assess the degree of achievement of an organization's objectives (e.g., to increase revenue), as well as to measure expected results within a business process (e.g., number of products sold). Strategic indicators are calculated from measures of interest collected from various sources and integrated into a multidimensional scheme. The measures are often of a corporate nature (sales, costs, customers, etc.), are generated within the same company and have a well-defined structure. However, today, much of the strategic information relevant to an organization resides in external sources, mainly in social networks (Zhou et al., 2015) (Fan and Gordon, 2014). Unfortunately, there are few studies that establish the most appropriate external indicators for each domain and the way to calculate them from the data offered by social networks.

Today, traditional BI processes related to decision making are affected by trends in social media, the latter providing immediate user feedback on products and services. In turn, new types of businesses have proliferated in digital media, newspapers, blogs, as well as digital marketing departments, whose market value is determined by user interaction, influence and impact on social media; their growth cannot be measured using traditional performance indicators.

From the BI point of view, social data can also be treated as a multidimensional model that can be linked to corporate data to aid decision making. In this area, we can define a social indicator as a time metric that allows an organization to dynamically measure the impact of its activities on social networks and the Web. The challenge lies in defining good social indicators from a large volume of unstructured data from social networks.

Given the interest in analyzing social networks to improve business processes, many commercial tools have proliferated for the analysis and monitoring of metrics and indicators in social networks, mainly offering statistical summaries of the metrics offered by the APIs of the most popular social networks. Most of these tools are limited to very specific contexts and dimensions, and do not allow a true integration with corporate BI systems. Some research focuses on modelling solutions to very specific problems such as the analysis of feelings, clustering of events, user classification and identification of marketing campaigns on Twitter. Currently, the analysis of social networks is reaching a sufficient

degree of maturity to be approached from a more methodological point of view.

In this paper we present a framework for the definition, capture and monitoring of social indicators based on the multidimensional model. The main objective is to provide a framework to facilitate the analysis of the semi-structured data offered in streaming by APIs services of various social networks (e.g. Twitter), and then summarize them as social indicators that respond to specific organization goals.

The rest of the article is structured as follows. In Section 2 we review the work related to solutions for social analysis, define the context of the research and the requirements that our proposal must meet. Section 3 defines the analytical patterns to be taken into account to develop a system oriented BI analysis. Section 4 presents the framework for defining and monitoring social indicators. Section 5 describes the characteristics and ways of evaluating a social indicator. Section 6 presents a case study to validate the proposal. Finally, we conclude the article in Section 7.

2 CONTEXT AND REQUIREMENTS

Our approach integrates a broad spectrum of research problems that have been addressed independently in the literature or that solve very specific problems. In the revised literature we identified four main research approaches that allowed us to group the related work together: "Social analysis for BI", "Streaming text processing", "Modelling performance indicators" and "Collaborative networks for maintaining performance indicators". These works are discussed in Section 2.1. Then in Section 2.2, we briefly present the background of our proposal and the new requirements for defining dynamic indicators for social network analysis.

2.1 Related Work

Social Analysis for BI. Despite the great commercial interest in creating analytical techniques for social networking, there are few approaches in the literature that address the issue within the area of BI. Some pioneering work has recently been reviewed in (Berlanga and Nebot, 2015), and basically they establish a correlation between external entities (such as news or opinions) and internal entities (the facts to be analysed). Other work has focused on creating multidimensional models for the analysis of opinions expressed in social networks about a product or company (Berlanga et al., 2015) (García-Moya, 2016). Many approaches in the area directly create ad-hoc processes that measure some kind of indicator on a given topic in a social network, mainly topological (Wang et al., 2013), product (Yan et al., 2015) (Chae, 2015), or feeling (polarity) (Dai et al.,2015) (He et al., 2015).

Streaming Text Processing. In (Feng et al., 2015) and (Liu et al., 2013) authors propose a similar approach as ours. They model and process streams of texts extracted from Twitter in the form of a multidimensional cube with the "TextCube" and "StreamCube" frameworks respectively. The former presents an algorithm for the detection, clustering and ranking of events through Twitter hashtags. These events are stored in a stream cube and the dimensions are limited to location and time. The second article is a study of human behaviour based on the analysis of feelings by geo-localization. In both reviews the data is stored on disk, requiring large amounts of storage resources to maintain the large volume of data generated by social networks. Moreover, indicators are restricted to a small set of predefined dimensions and metrics.

Modelling Performance Indicators. In the area of formalization and evaluation of performance indicators in a company, (Barone et al., 2011) includes a series of techniques and algorithms to derive composite indicators based on the Business Intelligence Model. On the other hand (Popova and Sharpanskykh, 2011) proposes a formal framework for the modelling goals based on performance indicators and defined mechanisms to establish the fulfilment of objectives, allowing the evaluation of organizational performance. Both approaches make it easier to derive indicators, to discover the relationships between them and to know clearly what they allow for evaluation.

Collaborative Networks for Maintaining Performance Indicators. Nowadays, the creation of collaborative networks are key factors in achieving sustainable competitive advantages for companies. Semantic technologies are a powerful tool to provide a common layer for information exchange. In this sense (Diamantini et al., 2016) establishes a semantic framework for the formal definition and collaborative maintenance of a dictionary of performance indicators. A similar approach (Maté et al., 2017), propose an infrastructure for automatic derivation of

company indicators, setting up a common framework between business analysts and developers that links business strategies and data analysis. The above proposals focus on the formal definition of indicators and highlight the importance of keeping them linked to business objectives. These solutions are a reference framework for the formalization of indicators in a company, but for Social BI it is necessary to manage data of a different nature (unstructured, volatile and fast) from external sources (unlike the historical measurements stored in DW). As a result, techniques for deriving performance indicators cannot be applied directly to social indicators because they are dynamic, volatile and less predictable in their behaviour.

2.2 Background and Requirements of the Proposal

In this paper, we consider as social information all collective information produced by customers and consumers in a marketplace when participating in online social activities. We will also refer to data extracted from social information by analysis tools, such as sentiment data or opinion facts. The amount of data extracted is massive so social forums can be considered as Big Data.

A previous work to this research is the SLOD-BI infrastructure (Berlanga et al., 2015). SLOD-BI provides mechanisms and tools to collect, store and analyze social metrics based on data published by social networks users. From a scientific and technical point of view, the project proposes the combination of cognitive models with statistical language models, large open knowledge resources and multidimensional analytical models to define efficient methods of extraction and analysis of social information. This infrastructure follows the principles of the Linked Open Data (LOD) initiative.

In this paper we propose to extend the SLOD-BI infrastructure with new modules for the definition of dynamic indicators for social analysis. The new requirements are the following ones:

1. Definition of a dictionary of social indicators correlated to the objectives of the business to be modelled.
2. Due to the dynamic nature of the data, the solution must allow the definition and updating of multidimensional structures for the analysis context.
3. Construction of a real-time data cube from linked semantic data and defined dimensions. The modelling of social measures in form of cubes allows the calculation and exploration of indicators on different dimensions.
4. The cube will keep only current information contained within given time windows. The information generated in social networks is constantly changing in function of new topics and trends that arise and disappear very quickly, so the most valuable social data that must be kept are the most current.
5. When an indicator is defined or updated, a new cube of social measures will be generated for it and its population will start from zero.

3 ANALYTICAL PATTERNS FOR SOCIAL BI

The main BI patterns identified for conducting the social analysis are summarized in Figure 1 (Berlanga et al., 2015). The links represent the relationship between the social data and the corporate data. The analysis patterns on the corporate data side correspond to the traditional models of a typical DW. While the patterns represented alongside the social data represent the multidimensional structures for social analysis. Facts are labelled with "F", dimensions with "D" and their levels "L". Facts directly involved in Social BI are: Opinion, Post and Social Facts.

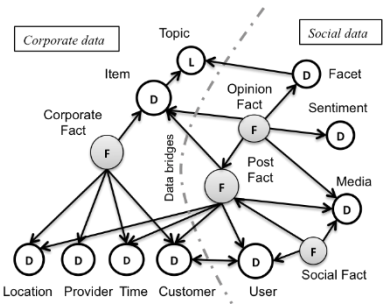


Figure 1: SLOD-BI Analysis Patterns.

Opinion Facts are observations based on types of feelings (e.g. positive or negative) expressed by users concerning specific facets (e.g. design) of an item of interest (e.g. a car brand). Post Facts are observations on the data of a particular post (e.g. reviewer, item reviewed, date reviewed), which may be related to a series of Opinion facts. Social Facts provide relevant information about users and their opinions in the context of the community to which they belong (Berlanga et al., 2015). The large volume of data generated around these patterns makes it difficult to interpret them in a timely manner, so it is necessary to define accurate aggregate mechanisms with different granularity in terms of space and time, as

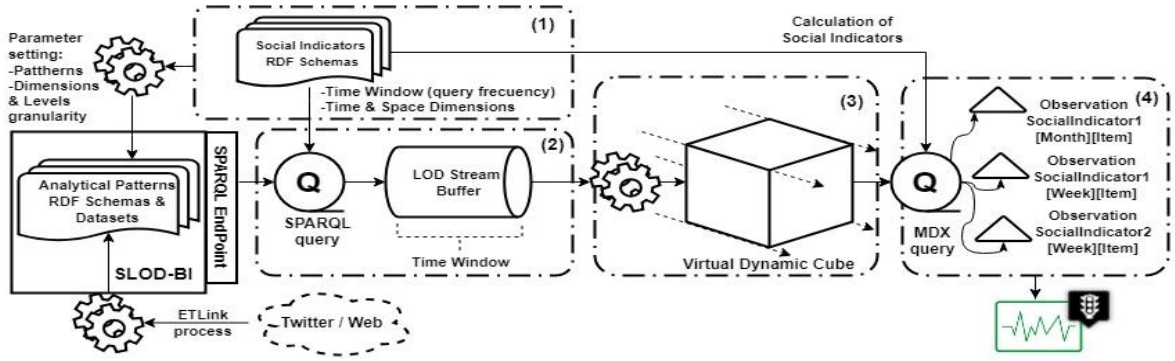


Figure 2: New Framework based on SLOD-BI infrastructure.

well as indicators to consolidate them into useful information. For this purpose we introduce a new high-level pattern: the social indicator, whose values will be dynamically derived from measures of the social patterns described above.

One of the main objectives of the infrastructure is to facilitate data integration by defining data bridges between corporate elements and social data (shown in the figure as dashed lines). Data bridges represent the process that allow to perform analysis operations that combine corporate and social data.

4 PROPOSED FRAMEWORK

The objective of the proposed framework is to allow the definition and derivation of social indicators for the analysis of social networks in streaming. The aim is to load the value of each social indicator into the strategic business model in order to help in the decision-making process. A social indicator is a new data pattern that is part of a layer that is above the SLOD-BI data infrastructure. Seen from top to down, the definition of a social indicator determines which data will be captured from social networks and how often they will be collected.

The proposed framework is summarized in Figure 2. It extends the previous SLOD-BI infrastructure with four new modules, namely: specification of the social indicators and dimensions of analysis (1), querying linked data patterns (2), construction of the Virtual Dynamic Cube (VDC) (3) and estimation of the social indicators facts (4).

Specification of Social Indicators and Dimensions. Key Performance Indicators (KPI) are typically expressed in technical terms using languages like MDX (MultiDimensional eXpressions) or SQL. As most social data in SLOD-BI are expressed as RDF (Resource Description Framework) triplets, we use OWL for describing social indicators formulas and dimensions. A social indicator is defined primarily by its name, formula and dimensions of analysis. The formula of a social indicator can be composed of social metrics and/or other indicators. It is also necessary to establish the periodicity of data collection. In Section 5 we present the semantics and rules for modeling social indicators.

In the model we define two main categories of dimensions: "Space Dimension" and "Time Dimension". The Space dimensions represent the context of the social data retrieved, e.g. domain, topic, item, user, location, and so on. The granularity of the time dimension will be determined by the frequency with which the different observations must be collected. Figure 3 shows the different dimensions characterizing a social indicator, which can be organized in hierarchies of analysis.

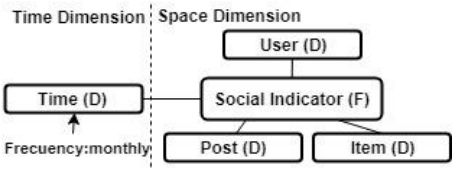


Figure 3: Analysis Dimensions.

Querying Linked Data Patterns. The SLOD-BI infrastructure must be parameterized with the metrics and dimensions (associated with each indicator) to be extracted from the social network. Social data will be captured during an ETLink process (Extraction, Transformation and publication of data in LOD). The SLOD-BI data service layer allows us to query the datasets through a SPAQRL endpoint. For each defined social indicator, a continuous query is defined, specifying its metrics, dimensions and time window (query periodicity). The result of the process is a stream buffer of linked facts for each social

indicator, which comprises the last date range associated with its time window.

Optionally, this output can be semantically enriched by adding new attributes extracted from the data itself. For example, using NLP techniques we can classify the texts of post facts in spam or not spam.

VDC Construction. Streamed linked social data will be transformed into a new multidimensional scheme that we call VDC inspired by the traditional OLAP data cube. Unlike traditional DWs where facts and measures are historically stored on disk, in our proposal the dimensional structures will be modelled "virtually" (they do not exist physically nor they are stored on disk, they are generated and processed on the fly). The "virtualized" data will materialize from the stream in the appropriate buffer and will have a temporary character. Measures or events must be generated periodically and their availability will be determined by the specified time window (e.g. last month) or by a number of previous observations (e.g. last 10 observations). To transform linked data into a multidimensional model there are several methods in the literature that can be applied (Nebot and Berlanga, 2016).

Dynamically generated data can be connected to external systems, such as: Exploration Tools, Predictive Models, Corporate DW or a Decision Support System.

Social Indicators and Facts Estimation. The numerical value of an indicator corresponds to an observation determined by the dimensions of the indicator and the observation date. The process of calculating a social indicator begins with an MDX query for the selection and aggregation of measures from the dynamic cube, and ends with the evaluation of its formula. Resulting values can be displayed in real time on a dashboard or a balanced scorecard. Optionally observations can be stored in a datawarehouse for historical analysis.

The indicators will exhibit a dynamic behaviour since their multidimensional structures may vary over time (e.g., adding or eliminating dimensions or measures). In this case, resulting observations could have different dimensional structures that must be taken into account when storing them.

5 MODELLING AND DERIVING SOCIAL INDICATORS

Similar to a KPI, a social indicator is defined by a mathematical expression or a specific value. Its basic properties are: name, definition, measuring objective, calculation periodicity, associated dimensions, unit of measurement, aggregate function, weight (importance), threshold, best and worst expected value. These last three properties will allow us to create visual alerts about the observations.

In this Section we propose an ontology to model social indicators. This extension corresponds to a high-level ontology within the SLOD-BI data schema. Figure 4 shows the main classes of the social indicators ontology. Table 1 shows the main OWL properties of the more general class "SocialIndicator". Letter “C” indicates the cardinality of the property.

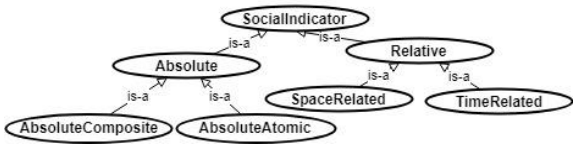


Figure 4: Class hierarchy for social indicators.

Table 1: Main properties of the SocialIndicator class.

Class	C	Property	Range
Social Indicator	>0	hasDimension	Item, User, Post
	=1	hasTime	Time
	=1	hasAggFunction	Sum,Avg,Max

A social indicator can be composite or atomic depending on the way it is calculated. In our model the atomic indicators are those that do not need a formula to be calculated, as their values are directly obtained from facts of SLOD-BI (e.g. number of post likes). On the other hand, the calculation of a composite indicator will depend on other predefined indicators.

It is important to clearly differentiate between two types of indicators that we often find in BI and we formalize in this study: absolute and relative indicators.

An “Absolute” indicator represents a numerical amount collected at a given time. This type of indicator can be either atomic or compound. An “Atomic” indicator represents a concrete measure directly obtained from the social network (e.g. number likes). On the other hand, a “AbsoluteComposite” indicator can be expressed as a mathematical expression whose arguments correspond to other “Absolute” indicators, either atomic or compound. The component indicators must have the same dimensional structures. Table 2 shows the properties and ranges that define the classes derived from the “Absolute” indicators.

Table 2: Main properties of Absolute Indicators Classes.

Class	C	Property	Range
Absolute Atomic	=1	hasMetric	SocialMetric
Absolute Composite	=1	hasBinary Operator	Binary Operator
Binary Operator	=1	hasMath Operator	Plus, Minus, Product
	=1	hasArgument1 hasArgument2	Absolute Indicator

Relative indicators are composite indicators whose values correspond to a ratio between two absolute indicators separated either in time or space. In case of space-related indicators, their calculation consists of a proportion (division). "Time Related" indicators imply a subtraction operation. Table 3 shows the main properties of the "Relative Indicator" class.

Table 3: Main properties of Relative Indicators Classes.

Class	C	Property	Range
Relative Indicator	=1	hasBinary Operator	BinaryOperator
Binary Operator	=1	hasMathOp	Minus, Division
	=1	hasArgument1 hasArgument2	AbsoluteIndicator

The class "SpaceRelated" is differentiated by the constraint: given two absolute indicators involved in the formula, the analysis dimension "A" of the first indicator must be a subset of the analysis dimension "B" of the second indicator ($A \subseteq B$).

The class "TimeRelated" is defined by the following constraint: given two absolute indicators involved in the formula, the time dimension "T1" of the first indicator must be disjoint from the time dimension "T2" of the second indicator ($T1 \neq T2$) and in turn must be structurally equivalent.

As examples, definitions 1 and 2 represent the social indicators Likes and Interactions respectively, while Figure 5 shows the properties of the Engagement indicator.

$$\text{Likes} \equiv \text{hasMetric.LikeMetric} \cap \text{hasDimension.Item} \cap \text{hasAggregationFunction.SUM} \quad (1)$$

$$\text{Interactions} \equiv \text{hasBinaryOperator.}(\text{hasMathOperator.SUM} \cap \text{hasArgument1.Likes} \cap \text{hasArgument2.Retweets}) \quad (2)$$

In the previous formulas we assume that all restrictions are functional (= 1).

Figure 5: Example of Engagement social indicator.

6 EXPERIMENTAL STUDY

With the purpose of validating the proposed framework to derive dynamic indicators, we have developed a prototype to address a use case in the car domain.

6.1 Case Study: Social Analysis in the Car Domain

The fundamental objective of any car rental company is to provide its customers with quality services and achieve effective sales. In addition to the traditional analytical queries that involve corporate data, there is a need to have a deeper insight of the business marketing processes in real time in order to react more efficiently. For a successful analytical experience, the company must specify the most important domains of analysis with the items (products or services) to be monitored.

In the context of the use case, the goal is to study the popularity of different car brands by tracking the "user's Engagement" in a given period.

6.2 Implementation of a Prototype

To populate the VDCs with real data we use a dataset of 2,625.186 tweets crawled using Twitter's streaming API from November 2014 to February 2017.

The developed workflow is based on the model explained in Section 4. The social indicators defined are: Engagement, Interactions (described in Section 5) and onDomain tweets (number of posts about the car brand). The metrics and dimensions that define each indicator are the input parameters for the SLOD-BI infrastructure to populate its datasets.

Once SLOD-BI is configured for the car rental domain, the sentiment data can be consumed via the data service layer to produce the required data. Table 4 shows the workflow of the implemented process, the operators involved and their corresponding input/output data.

Table 4: Proposed workflow and operator types.

Operator Type	Input	Output
QuerySparql	Sparql query	RDFStream
Continuously extracts the union/intersection of RDF social data bounded by the dimensions and time window of the social indicator.		
DataEnrichment	RDFStream	RDFStream
Optionally, predictive models can be applied to the output data (e.g. determine whether or not a post is spam) and the RDF can be enriched with new predicates.		
VDC construction	RDFStream	MDXStream
VDC construction from streamed linked data.		
IndicatorCalculation	QueryMDX	Value
Evaluates the mathematical operations in the MDX query. The query frequency is determined by indicator.		

In our simulation, the indicator facts table was saved in a CSV file for viewing it in the Tableau tool.

6.3 Visualization and Analysis

Below are a series of examples of interesting analytical queries to monitor the interest that the company arouses in the social network users.

The analyst wants to check if a Twitter marketing campaign was effective. For this purpose, it is necessary to analyze the response of users in the corresponding period through the defined social indicators. Figures 6 and 7 show the values of the Engagement indicator for different cars brands for the whole period. The first one shows the result for all users of the dataset (spammers included), while the second one shows only the values for non-spammers users in which we check a more linear result. For this segmentation we use the entire dataset for train a Spam classifier with a Linear SVM. The classifier was implemented in Python with the Anaconda framework (Pandas and Scikit-Learn packages). After applying the Spam Classifier, the number of events is reduced by around 40%.

The analyst wants to check the impact on Volkswagen car rentals, after the controversy generated when the Environmental Protection Agency revealed in September 2015 that the manufacturer had manipulated the emissions detection software.

Figure 8 shows the result of the onDomain tweets and Engagement indicators for different brands of interest during the period of dispute. The graph shows clearly the high impact of Volkswagen brand posts.

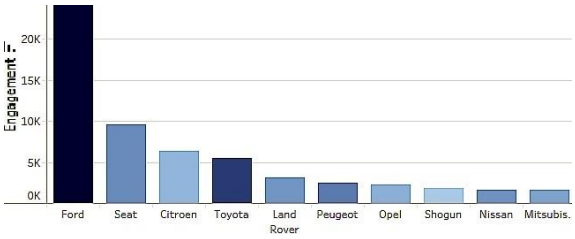


Figure 6: Engagement indicator with all users included.

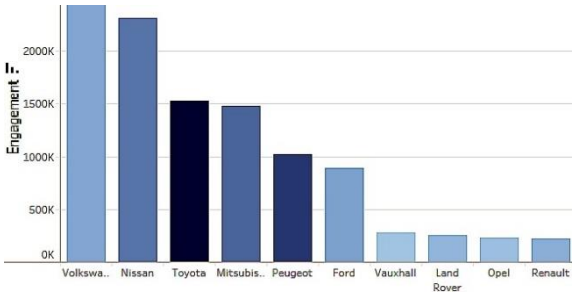


Figure 7: Engagement indicator without spammer's users.

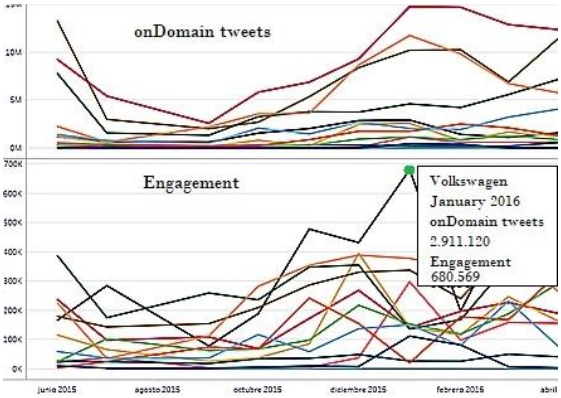


Figure 8: Engagement and onDomain tweets indicators.

7 CONCLUSIONS

In this article, a novel approach has been presented for the definition and monitoring of social indicators on the linked and open data infrastructure called SLOD-BI. The proposal offers the possibility of exploring the measures captured from social networks interactively over different multidimensional contexts and in real time.

We propose a framework that makes use of the principles of LOD data to define and publish as semantic data the definitions of social indicators. On the other hand, the indicator measurements are calculated on the fly from a linked social data stream modelled like an OLAP cube, but keeping only the

most recent information. It is important to highlight the dynamism of the cube as it supports the continuous inclusion of new measures and dimensions.

Among the main benefits of this framework is the fact that the indicators are directly linked to the social measures, so that it is possible to easily identify the origin of the values of these indicators. On the other hand, the fact that the indicators are also semantic data, makes it possible to apply validation techniques during their definition and derivation.

As future work will be studied the automatic creation of descriptions and queries associated with the calculation of social indicators, as well as the discovery of appropriate metrics to evaluate strategic objectives of the organization. Due to the dynamism of the cubes, the volume and fluctuating character of the data, makes it impracticable to store historical data, so it is necessary to establish the appropriate mechanisms to find the right time window to apply predictive algorithms and compare measurement trends.

ACKNOWLEDGEMENTS

This work has been financed by the Ministry of Economy and Trade with the project of the National R&D Plan with contract number TIN2017-88805-R. We also have the support of the Universitat Jaume I pre-doctoral scholarship programme (PREDOC/2017/28).

REFERENCES

Barone, D., Jiang, L., Amyot, D. and Mylopoulos, J., 2011. Composite Indicators for Business Intelligence. *Conference on Conceptual Modeling ER 2011. Lecture Notes in Computer Science*, 6998, pp. 448–458.

Berlanga, R. and Nebot, V., 2015. Context-Aware Business Intelligence. *Business Intelligence. Lecture Notes in Business Information Processing*, 253, pp. 87-110.

Berlanga, R., García-Moya, L., Nebot, V., Aramburu, M., Sanz, I. and Llidó, D., 2015. SLOD-BI: An Open Data Infrastructure for Enabling Social Business Intelligence. *International Journal on Data Warehousing and Data Mining*, 11(4), pp. 1-28.

Chae, B. K., 2015. Insights from hashtag# supplychain and Twitter analytics: Considering Twitter and Twitter data for supply chain practice and research. *International Journal of Production Economics*, 165, pp. 247-259.

Dai, W., Han, D., Dai, Y. and Xu, D., 2015. Emotion recognition and affective computing on vocal social media. *Information & Management*, 52(7), pp. 777-788.

Diamantini, C., Potena, D. and Storti, E., 2016. SemPI: A semantic framework for the collaborative construction and maintenance of a shared dictionary of performance indicators. *Future Generation Comp. Syst.*, 54, pp. 352-365.

Fan, W. and Gordon, M. D., 2014. The Power of Social Media Analytics. *Communications of the ACM*, 57(6), pp. 74-81.

Feng, W., Zhang, C., Zhang, W., Han, J., Wang, J., Aggarwal, C. and Huang, J., 2015. STREAMCUBE: Hierarchical spatio-temporal hashtag clustering for event exploration over the Twitter stream. *IEEE 31st International Conference on Data Engineering*, pp. 1561-1572.

García-Moya, L., 2016. Modeling and analyzing opinions from customer reviews. Tesis Doctoral, Departamento de Lenguajes y Sistemas Informáticos, Universitat Jaume I, Castellón.

He, W., Wu, H., Yan, G., Akula, V. and Shen, J., 2015. A novel social media competitive analytics framework with sentiment benchmarks. *Information & Management*, 52(7), pp. 801-812.

Liu, X., Tang, K., Hancock, J., Han, J., Song, M., Xu, R., and Pokorny, B., 2013. A Text Cube Approach to Human, Social and Cultural Behavior in the Twitter Stream. *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, pp. 321-330.

Maté, A., Trujillo, J. and Mylopoulosb, J., 2017. Specification and derivation of key performance indicators for business analytics: A semantic approach. *Data & Knowledge Engineering journal*, pp. 30–49.

Nebot, V. and Berlanga, R., 2016. Statistically-driven generation of multidimensional analytical schemas from linked data. *Knowledge-Based Systems*, 110, pp. 15-29.

Popova, V. and Sharpanskykh, A., 2011. Formal modelling of organisational goals based on performance indicators. *Data & Knowledge Engineering*, 70(4), pp. 335-364.

Wang, G. A., Jiao, J., Abrahams, A. S., Fan, W. and Zhang, Z., 2013. ExpertRank: A topic-aware expert finding algorithm for online knowledge communities. *Decision Support Systems*, 54(3), pp. 1442-1451.

Yan, Z., Xing, M., Zhang, D. and Ma, B., 2015. EXPRS: An extended pagerank method for product feature extraction from online consumer reviews. *Information & Management*, 52(7), pp. 850-858.

Zhou, M., Lei, L., Wang, J., Fan, W. and Wang, A. G., 2015. Social Media Adoption and Corporate Disclosure. *Journal of Information Systems*, 29(2), pp. 23-50.

Chapter 3

Modelling Analytical Streams for Social Business Intelligence

Publication

Lanza Cruz, Indira Lázara; Berlanga, Rafael; Aramburu, María José. Modeling Analytical Streams for Social Business Intelligence. Informatics 2018, 5, 33. https://doi.org/10.3390/informatics5030033. (Q1)

informatics

MDPI

Article

Modeling Analytical Streams for Social Business Intelligence

Indira Lanza-Cruz [1], Rafael Berlanga [1,*] and María José Aramburu [2]

[1] Department de Llenguatges i Sistemes Informàtics, Universitat Jaume I, 12071 Castelló de la Plana, Spain; lanza@uji.es

[2] Department de'Enginyeria i Ciència dels Computadors, Universitat Jaume I, 12071 Castelló de la Plana, Spain; aramburu@uji.es

* Correspondence: berlanga@uji.es; Tel.: +34-964-72-8367

Received: 25 June 2018; Accepted: 23 July 2018; Published: 1 August 2018

Abstract: Social Business Intelligence (SBI) enables companies to capture strategic information from public social networks. Contrary to traditional Business Intelligence (BI), SBI has to face the high dynamicity of both the social network's contents and the company's analytical requests, as well as the enormous amount of noisy data. Effective exploitation of these continuous sources of data requires efficient processing of the streamed data to be semantically shaped into insightful facts. In this paper, we propose a multidimensional formalism to represent and evaluate social indicators directly from fact streams derived in turn from social network data. This formalism relies on two main aspects: the semantic representation of facts via Linked Open Data and the support of OLAP-like multidimensional analysis models. Contrary to traditional BI formalisms, we start the process by modeling the required social indicators according to the strategic goals of the company. From these specifications, all the required fact streams are modeled and deployed to trace the indicators. The main advantages of this approach are the easy definition of on-demand social indicators, and the treatment of changing dimensions and metrics through streamed facts. We demonstrate its usefulness by introducing a real scenario user case in the automotive sector.

Keywords: Social Business Intelligence; data streaming models; linked data

1. Introduction

The main objective of Business Intelligence (BI) is extracting strategic knowledge from the information provided by different data sources to help during the decision-making process and achieve the strategic goals of a company. The processing and analysis of massive data oriented to BI has evolved in recent years. Traditionally, the most commonly used approaches have combined data warehouse (DW), online analytical processing (OLAP), and multidimensional (MD) technologies [1], on very specific scenarios, making use of static and well-structured data sources of corporate nature, being all the information fully materialized and periodically processed in batch mode for future analysis. More recently, new technologies for Exploratory OLAP have been introduced, aimed at exploiting semi-structured external data sources (e.g., XML, RDF) for the discovery and acquisition of relevant data that can be combined with corporate data for decision making process.

Today, BI processes related to decision making are affected by trends in social media, the latter providing immediate user feedback on products and services. Social networks are a fundamental part of the information ecosystem, social media platforms have achieved an unprecedented reach for users, consumers, and companies providing a useful information channel for any professional environment. For the above reasons, there has been a growing interest in the development of solutions from the commercial and scientific perspectives. However, there are peculiarities that do not allow for the

direct adaptation of traditional BI techniques because the social data to be analyzed are considered Big Data (i.e., big volume, unbounded, heterogeneity, semi and unstructured data, volatility, and speed or streaming).

The development of new methods and architectures for Big Data analysis has also evolved. Currently, we can clearly differentiate two trends, namely: Traditional and Streaming analytics architectures for Big Data. The first is the most widely used in the literature, integrating various Big Data sources into a raw data repository under a multidimensional scheme (DW/OLAP). All the information is stored in a historical repository, although sometimes the dynamism of the data makes it unnecessary wasting resources. Data processing is done in batch, which causes late alerts and a delay in the decision making of a BI system.

A newer approach and in accordance with the current needs for Big Data processing, focuses more on the speed and immediacy of information, processing data in streaming and, depending on the model designed, incorporating batch analysis processes to obtain knowledge models. In this way, they can offer fresh analysis data and enriched information from models. Only the information needed for the knowledge models are stored, optimizing memory usage.

However, these approaches are still in development and the literature has addressed approaches aimed at solving very specific problems, combining technologies such as Stream Processing Engines (SPE) together with OLAP systems. The main difficulties encountered are the deployment of independently created technologies and connecting them through ad-hoc processes, which can lead to unnecessary replication of information and performance degradation. The analysis dimensions are treated in a static way, when the nature of Big Data has a dynamic and unlimited multidimensionality. They are not integral solutions neither extensible nor scalable.

Thus, the goal of this paper is to propose a generic architecture for the analysis of Big Data, which allows the processing of dynamic multidimensional data in streaming. The main contributions of this work can be summarized as follows:

- We present a comprehensive revision of the main methods proposed for Social Business Intelligence.
- We propose a streaming architecture specially aimed at Social Business Intelligence.
- We propose a new method to model analytical streams for Social Business Intelligence.

The rest of the article is organized as follows. In Section 2 we review the work related to solutions for social analysis, identifying the main methods and tasks of analysis addressed by the scientific literature. Section 3 presents the proposed architecture and Section 4 a prototype implementation with a use case based on the car sector. Finally, Section 5 presents the main conclusions.

2. Methods and Task for Social Analysis

The processing and analysis of Big Data is a field of research still very young. Recently several authors have proposed some architectures for streaming data processing, among them we can highlight the Streaming Analytics, Lambda, Kappa, and unified architectures [2]. So far, these proposals do not constitute standards since they have not been evaluated enough, nor validated on several domains. When selecting an architecture, it is essential to clearly know the uses cases that are to be modeled. With the aim of developing a generic architecture that covers a wide range of Big Data analysis tasks, we have developed a state of the art identifying the main tasks and the most used methodologies for processing them. Our review is focused on relevant articles that propose solutions for processing social data in streaming, especially those data offered by the API services of Twitter.

Table 1 summarizes the findings related to the analysis tasks. In the reviewed literature, we have identified a large set of analysis tasks that have been grouped into six main categories: sentiment analysis, user profiling, post profiling, event detection, user network interactions, and systems recommendations. Most of these tasks focus on the analysis of behavior patterns and features that can be coded and adopted by machine learning techniques for classification. Different kinds of features

classes are commonly used to capture orthogonal dimensions for the objects of analysis (e.g., users, posts, events), as summarized in Table 2.

Table 1. Analysis tasks for social data.

Category	Analysis Tasks	Features Class	References
Sentiment Analysis	Sentiment indicators. Communications analysis. User, groups, community, society. characterization Human, social and cultural behavior.	Post content	[3–6]
User Profiling	Author profiling, User classification. Inferring user properties (age, gender, income, education, etc.). Political orientation, ethnicity, and business fan detection. User interests identification.	User Metrics, Post content, Post metrics, Sentiment, Network	[7–10]
	Spammers, Bots detections, Promoters, Influencer detection.	All	[11–17]
Post Profiling	Campaigns, Topics, spam, meme, sarcasm, rumors, terrorism detection.	Post, Links, Bursts	
Event detection	Real-time events detection by location and time, events classification, protests and manifestations, detection of diseases and catastrophes, study of the displacement of people between cities. Real-time classification of trends.	Post content, Post metrics, Hashtag, Location, Time, Burst	[18–23]
Analysis of social network and Users interactions	Influence and correlation in social networks, social network dynamics, network, and node classification, detect influential nodes.	Graph Network, User metrics	[24]
Recommendations system	User, news, media recommendation.	User metrics, Post metrics, Time	[25]

Table 2. Classes of features used in different tasks for social analysis [14].

Class	Description
User metrics	The user features refer to the metadata related to a user account on social networks. You can include data such as geo-location, friend list, number of mentions, etc.
Post content and metrics	Post features can be divided into two main parts: text contents and Post meta-data. With the text it is possible to analyze its content and identify clues based on the linguistic characteristics making use of natural language processing algorithms. From the text it is also possible to extract links, hashtags or embed multimedia. On the other hand, the metadata refer to the records of user interactions with the post, such as the number of responses, retweets, likes or date of publication.
Network	The initial aim of the network analysis is to capture the basic perceptions of its macrostructure. At the micro level, network analysts focus on the importance of individual nodes. Network features capture various dimensions of information dissemination patterns. Statistical features can be extracted from retweets, mentions, and hashtag co-occurrences. From the global network, metrics such as number of nodes, number of edges, density, and diameter can be extracted; the main task includes node and edge classifications based on degree, inter-centricity, and proximity centrality. On the other hand, analyses are carried out to search for communities and to compare typified nodes [26].
Burst	A burst is a particular moment when the volume of tweets suddenly increases drastically. Burstiness is a spatio-temporal feature. We can measure how temporally bursty a keyword is at a location, and inversely in a concrete timing we can measure spatial burstiness [15,19].
Time	Time features capture the temporal patterns of posting and post interactions (e.g., replies, retweets, likes); for example, the average time between two consecutive publications.
Sentiment	Sentiment features are built using sentiment analysis algorithms from the posts content. It is also possible to use metrics such as ratings, emotion, and satisfaction scores.

Below we summarize the main methods for processing Big Data from Twitter. We divide existing proposals in two main methods for Intelligent Data Processing, namely: *Semantic Enrichment* and *Inductive Processing*, represented in Tables 3 and 4 respectively. Since some exhibit both elements we also present methods that mix ideas from these two main approaches and are also shown in Table 4.

Table 3. Methods for intelligent social analysis that use Semantic Enrichment.

Model Type	KFC	LOD	References	Tasks Types
Social Business Intelligence, Batch processing, OLAP based	Social Facts Ontology, ETLink, OLAP, Analytic Tool	RDF	[3,27,28]	Sentiment analysis. Entity extraction, keyword extraction, event, and topic detection.
	OLAP + ETL + Analytic Tools	No	[29,30]	
SMA [1], Cube modeling, Batch processing	OLAP	No	[23]	Spatio-temporal analysis of social media data. Detection of diseases and catastrophes. Study of people displacement.
Text Cubes Batch processing	OLAP	No	[4]	Sentiment analysis. Study human, social and cultural behavior.
RDF Streams. Streaming processing	R2RML mappings	RDF streams	[31]	Publish and share RDF streams on the Web.
	Streaming linking Data Server+ HTML5 browser	RDF streams	[32]	Sentiment analysis. Local events monitoring. Hasthags ranking.

[1] Social Media Analytics.

Table 4. Methods for intelligent social analysis that use Inductive Processing and Mixed.

Model Type	KFC [1]	LOD [2]	References	Tasks Types
SMI [3], Batch processing	ML framework	No	[8]	Political orientation, ethnicity, and business fan detection
SMI, Streaming processing	Online mode, Batch mode, Event Ranker	No	[19]	Real-time local event detection
StreamCUBE Batch processing Disk-based Storage	Spatial-temporal aggregation, Hashtag clustering, Event Ranker	No	[18]	Spatio-temporal hashtag clustering for event exploration
(Mixed) RDF Streams Streaming processing	DSMS [4], Abstracter DSMS, Deductive and Inductive reasoner	RDF Streaming OWL2-RL	[25]	User profiling and media recommendations using deductive and inductive reasoning.

[1] Key Framework Components. [2] Linked Open Data. [3] Social Media Intelligence. [4] Selector data stream management systems.

Providing *Semantic Enrichment* refers to extract the semantic information hidden in the post texts and metadata. *Semantic enrichment* is achieved as the combination of different (and possibly alternative) techniques [29]: crawler meta-data, information retrieval, crawler sentiment, NLP (Natural Language Processing) analysis, and domain experts.

The methods revised in the literature based on *Inductive Processing* make use of different techniques, see Figure 1, for solving more complex tasks that require more in-depth learning such as real-time event detection, user and post profiling, marketing campaigns, spammers or bots detection, as well as development of referral systems.

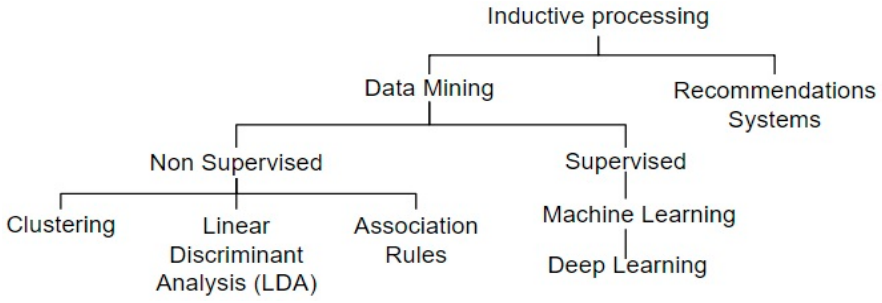


Figure 1. Inductive processing techniques used for intelligent social analysis.

We have also identified some important dimensions of analysis that group and distinguish the work reviewed, namely: model type, whether the system does streaming processing, Key Framework Components (KFC), and if they make use of Linked Open Data (LOD) technologies. The last columns at each table list the specific analysis tasks that are developed in each proposed method.

Although, in Tables 3 and 4 we have pointed out several types of models, three of them are worth mentioning because they really represent the evolution of the technological solutions to process Big Data: first social media analytics (SMA), then the proposals towards social media intelligence until reaching social media BI.

The work related to SMA mainly proposes tools and frameworks to capture, monitor, summarize, and visualize social metrics and sentiment analysis, usually aimed at solving specific use cases. It is important to highlight that the revised works have evolved towards more efficient mechanisms to organize information: social data are modeled within a multidimensional scheme, they organize linguistic features and sentimental indicators in a cube data model, facilitating query and visualization from various perspectives. These approaches usually integrate multiple engines to achieve OLAP over evolving data streams, calculations, and aggregations are done in batch so data analysis is deferred.

On the other hand, the goal of social media intelligence is to obtain actionable information from social media in context-rich application environments, to develop corresponding decision support frameworks, and to provide architectural designs and solution frameworks for applications that can benefit from the intrinsic knowledge of social conversations [33].

In this study the solutions reviewed in this study mainly make use of inductive processes in batch, so it is not possible to obtain insights in real-time. The most frequently used analysis tasks are event detection, user and post profiling and very simple recommendation systems. Research on social media intelligence is still at an early stage of development, despite the growing interest of companies and different professional areas that could benefit from such studies.

In the field of Social BI there are very few approaches in the literature. In this sense we highlight the contributions of [3,29], who propose frameworks to exploit BI by integrating social data. In the work [29], the authors propose the development of an OLAP architecture for social media analysis, while [3] offers a semantic LOD-based data infrastructure for the capture, processing, analysis, and publication of social data in RDF, so that it can be easily consumed by BI systems. Both propose a traditional approach since their architectures do not process the data in streaming (real-time), in turn materialize all the information in storages, so they are oriented to cases of high latency use.

In the literature reviewed, few solutions actually do streaming processing, analysis tasks are mainly oriented towards event detection [19] and recommendation systems [25]. It should be noted that both works make use of Semantic Web (SW) technologies to structurally enrich the processed data, allowing for reasoning tasks on them and facilitating the link with external sources. They take LOD technology to a higher level by proposing to share semantic information in the form of RDF streams. However, systematic research and well evaluated results are still lacking.

The work in [34] proposes an extended lambda architecture for Big Data processing that includes a semantic data processing layer and, similar to our proposal, it establishes mechanisms to semantically enrich raw data with metadata from various sources. Batch data processing is carried out in parallel with streaming processing. At the speed layer, events are validated and routed for either real-time or batch processing. Compared to the previous solution, our architecture allows mutual retrofitting between two parallel stages (see Section 3) for the improvement of each of their processes. It adds a layer for the inclusion of intelligent data processing algorithms implemented by the data scientist. We model stream types and stream workflows to facilitate data processing and Social BI oriented analysis. In addition, queries only need to search in a single service location rather than in batch and real-time views.

Social media data are dynamic streams whose volume is growing rapidly so it is necessary to create efficient mechanisms for their processing. Most of the current proposals focus on social media analysis aimed at solving problems in a closed context and do not solve the problem of the dynamism,

variety, and speed of social data. In this sense it is still necessary to establish mechanisms to discover and add new dimensions of analysis and new types of facts dynamically. It is necessary to create mechanisms to classify data according to the context. It is necessary to solve issues such as semantic consistency check, conflict detection, absence of structures and the difficulty of integrating different types of data and components in the same infrastructure.

To offer a solution to the previous problem situation, in this article we propose a unified, generic architecture that facilitates the deployment of an intelligent social analysis system of easy integration with BI systems. In this sense the use of SW technologies to model data streams as a multidimensional model facilitates the integration of structures and data between systems.

Following the guidelines of [35], we will now set out the requirements that the proposed architecture must meet:

- Keep the data moving. To support real-time stream processing, messages must be processed in streaming. Due to the low latency of social data, the system must be able to process data "on the fly" and avoid costly storage operations.
- Support for ad-hoc stream queries to filter out events of interest and perform real-time analysis. Support for high-level query languages for continuous results and set up primitives and operators to manage common stream properties (such as data window size and calculation frequency). Some streaming languages include StreamSQL and C-SPARQL.
- Establish mechanisms for handling stream imperfections. In real applications, streaming data may arrive out of order, with some delay, with missing information or arrive in the wrong format. The system must be able to identify the type of error and offer automatic solutions for each case.
- It must be robust and fault tolerant. The first property is related to the possibility of dealing with execution errors and erroneous inputs, in turn guaranteeing outputs in accordance with the expected results. Furthermore, it is necessary to guarantee the availability and security of the data, so that if any module has a failure, the system can continue working (to face this it is convenient to keep processes in the background that often synchronize states with primary processes).
- Allow the integration of stored data with streaming data. A streaming architecture must support integrating data from two basic states: a long-term stage for keeping and batch processing historical information and a short-term stage for generating data in streaming. In many scenarios, it is necessary to compare a current state with past states in order to obtain better insights from the data (e.g., for machine learning tasks). It is therefore necessary to efficiently manage the storage and access of previous states. On the other hand, it is not always necessary to store historical data forever, but it is recommended to establish a time window that maintains the last period of interest so that in case of process failure it is possible to recalculate all data from the latest historical data, thus supporting fault tolerance.
- The system must be partitionable and scalable automatically. That is, it must be able to automatically balance process overload, distributing processes in threads transparently to the user.
- High speed processing and response. It must support high performance for a large volume of very low latency streaming data.
- Finally, we have included integration with LOD technologies to semantically enrich the input and output of data. SW technologies enable the linking and exploration of external sources for the discovery and acquisition of relevant data (e.g., the discovery of new dimensions of analysis). On the other hand, it is also useful to enable the publication of streaming data in some standard for LOD so that it can be easily understood by external applications.

3. Proposed Architecture

In this section we describe the proposed architecture for analytical processing of Social BI according to the tasks and requirements discussed in previous sections.

First of all, we include two kinds of actors in the architecture, namely: data scientists and data analysts. The former is in charge of defining intelligent data processing tasks over the stream of data whereas the latter is in charge of defining over the resulting flow of measures the analytical tasks that trace the organizational goals and indicators. In this scenario, data analysts fetch requests to data scientists for inferring new perspectives from data (e.g., classifiers) and data scientists can demand new data streams to feed their analysis processes. The role of the data scientist is to define, develop, implement, and test inductive processing algorithms. As shown later, data analysts will define the analytical data streams as they are composed of analysis facts.

As Figure 2 shows, the proposed architecture allows for the reprocessing of analytics on historical data whenever a new logic needs to be applied. It roughly corresponds to a Kappa streaming architecture [2,36], which is an evolution of the Lambda architecture [37] in order to avoid the implementation of functionality twice for batch and real-time streaming. Our architecture basically consists of two streaming stages: a long-term stage for keeping recent historical information as a data stream of long duration and a short-term stage with some workflows for generating the required analysis data in real-time. In our scenario, data scientists get samples with training data for their algorithms from the long-term stage and test them over the data in the short-term stage. Data analysts usually consume data from the short-term stage through the serving layer. The long-term stage is also used to recover data when the stream workflow needs to be reconfigured or updated. In this case, whenever the short-term stage stops generating data and once it starts over, it needs to re-compute all the income data stored at the long-term stage.

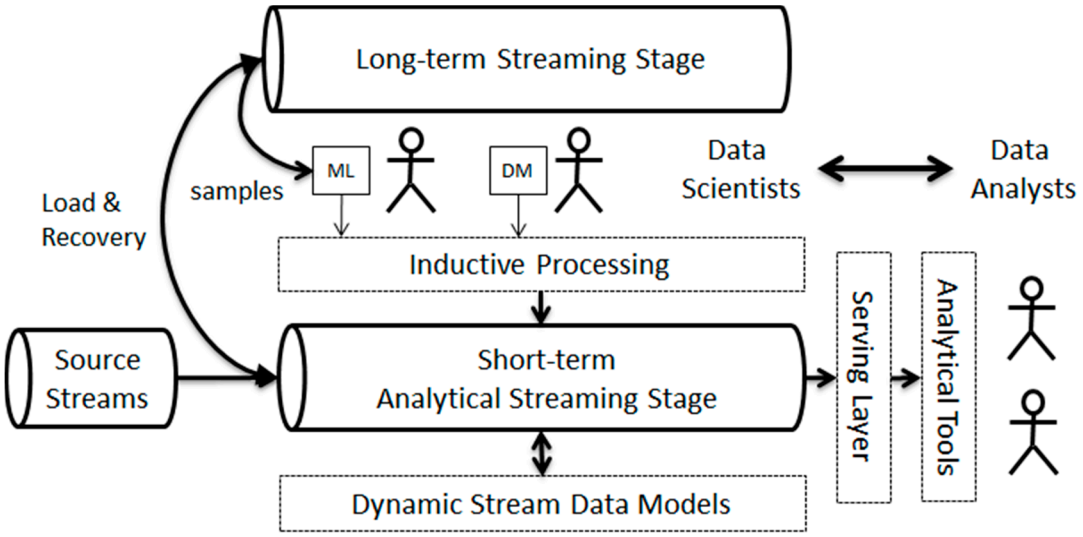


Figure 2. The proposed architecture for Dynamic Social Business Intelligence.

In our proposal, source data streams are semantically enriched through a series of vocabularies to produce useful data for analysis tasks. Similarly, all the produced and consumed data streams are factual and semantic, as they consist of facts and dimension data linked to the elements of the multidimensional analysis model. This allows the entire system to maintain the stream data models up-to-date with respect to the incoming data and analysis models. Any differences between them will cause the updating and reconfiguration of the corresponding workflows.

In the Lambda architecture the batch layer is a repository that stores the raw data as it arrives for processing by batch processes that will occur on some interval and will be long-lived. The stored data will be processed by iterative algorithms and transformed according to the needs of analysis [34]. While the speed layer is used to compute the real-time views that complement the batch views in the serving layer. The batch layer is mainly used to produce results without stringent latency constraints and to rectify the errors generated in the speed layer. With the Kappa architecture, when required,

the batch layer also feeds the real-time layer, so that both real-time and batch processing is executed by the same real-time data processing module.

In contrast, in the proposed architecture, data is always processed as a continuous stream. As in the Kappa architecture, the idea is to handle both real-time data processing and reprocessing in a single stream processing engine. In contrast to the Lambda and Kappa architectures, the long-term stage is a temporary data buffer that can be used for two purposes. First as continuous input for Machine Learning algorithms, whose outputs are used during the PRE and POST processes of the short-term stage. And secondly, it is also used for data recovery in case of system failure or stream workflow reconfiguration. Short-term stage corresponds to the same speed-layer as the Lambda and Kappa architectures.

The proposed architecture offers many advantages for each type of actor involved. It first clearly establishes the flow and channels for information exchange with the system and between actors. Secondly, it greatly accelerates communication and information exchange, as it reduces the time for data preparation and data collection thanks to the PRE and POST processes (see Section 3.1). Thirdly, it facilitates the work of the actors, allowing them to obtain results that are tailored to their needs and on demand. Many of the processes are often performed manually for the preparation and integration of data, our architecture allows to execute them automatically or semi-automatically.

In the next section we discuss how the elements of these architecture are modeled to build the complex workflows that fulfill Social BI tasks.

3.1. Stream Modeling

We basically distinguish two groups of data streams in the architecture: source streams and fact streams. Source streams are in turn divided into data streams and linked data streams. Data streams are directly connected to the sources of incoming social data (e.g., posts, user updates, etc.) whereas linked data streams leverage semantically the enrichment of incoming data to properly shape them for analysis. Fact streams are placed at the short-term stage with the main purpose of producing ready-to-use analytical data. In other words, fact streams process and aggregates source data to provide useful insights to data analysts.

Fact streams are fed by source data streams. For this purpose, they need to transform incoming data into multidimensional facts. We call these transformations ETLink processes. The name ETLink comes from the traditional Extract/Transform/Load phases, but instead of loading the transformed data, ETLink produces linked data graphs on the fly [3]. This process needs domain ontologies and controlled vocabularies to annotate, normalize, and link data. As these semantic sources can also evolve along time, we would need a third type of streams: the linked data streams.

Linked data streams are directly connected to Linked Open Data (LOD) endpoints, accessed via SPARQL queries [38]. Although many LOD sources are static, there is a great trend towards dynamic LOD (e.g., Live DBPedia [39] or Live BabelNet [40]), which ensures the provision of fresh semantic data. If a LOD source feeds some fact stream within the short-term stage, it must be transformed into a multidimensional model. We call this process a MD-star process. We can find several methods in the literature aimed at fitting linked data into multidimensional models [28,41], which can be considered to implement MD-star processes.

In Figure 3 we show the graphical representation of the previous stream types for the proposed architecture. We model each stream with three main components: a query Q and two optional processes (PRE and POST). Incoming data can be pre-processed by a PRE-operation, then Q is executed on the resulted data, and eventually a POST operation can be applied to produce the output data. The previously defined ETLink and MD-star processes are two types of POST operations.

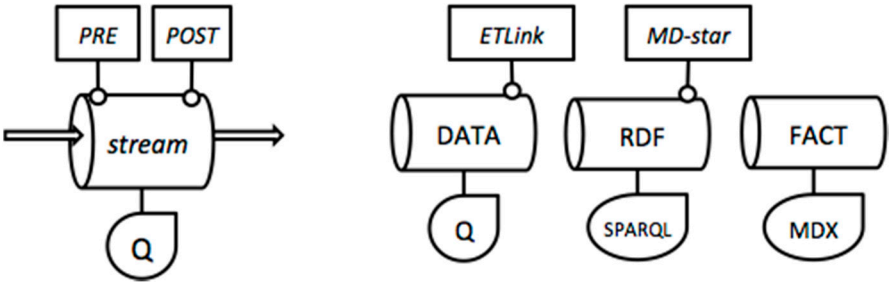


Figure 3. Stream graphical models for the proposed architecture.

In the case of fact streams, PRE and POST operations can only add new measures and dimension attributes before (PRE) or after (POST) the analytical query (MDX) is executed. An example of PRE-process would calculate the tweet probability to be in the domain. POST processes make inferences and calculations from aggregated data, for example, to determine the profile of a user after aggregating the metrics of the related posts. As shown in Figure 4, some processes can be fed by other streams to get fresh data required by their algorithms. This is the case of continuous learning algorithms like k-means, LDA and some machine learning methods like neural networks.

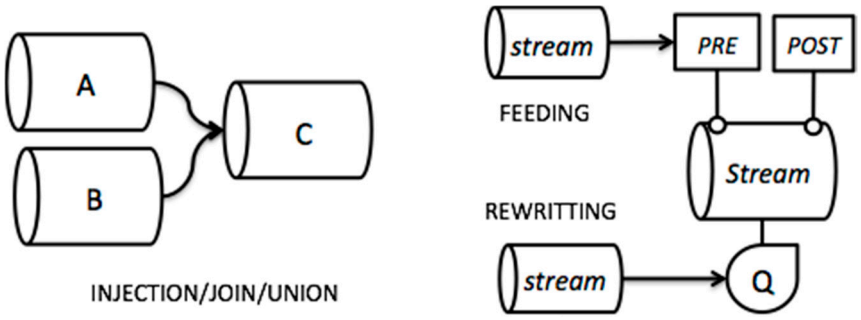


Figure 4. Modeling workflows of streams.

Additionally, any stream is characterized by its temporal behavior. Basically, each stream defines a time window and a slider. The time window indicates for how long data must be kept in the stream, and the slider indicates the periodicity or frequency of the output. For example, a time window of a week with a slider of one day means that the stream keeps the income data of the last seven days and generates an output every day. These two parameters depend on the analytical task at hand and can be derived from the specification of the MDX query. It must be pointed out that in this scenario, MDX queries must be tuned with respect to time dimensions by using relative tags like now. This is the same issue as for any continuous query language like C-SQL and C-SPARQL [42].

Modeling streams workflows with this approach has several advantages. The coherence between components can be checked before execution, and the resulting models can be dynamically updated as data is processed. For example, model updating will be necessary whenever either new dimensions attributes/members or new metrics appear in the source streams.

3.2. Multidimensional Coherence

Checking multidimensional coherence basically implies to infer two data schemas for each stream: the input data schema (IS) and the output data schema (OS). Notice that stream transformation process consists of three steps: (IS) PRE→Q→POST→(OS). From the stream definitions it is possible to infer how the IS is being transformed into the OS, and consequently analyze the whole processing of the workflow.

The IS of a stream will depend on the data streams it consumes. More specifically, the IS of a stream is the proper combination of the OS of the streams it consumes. As for fact streams, stream combination can have several interpretations according to the multidimensional paradigm:

The IS of the combined stream represents the injection of attributes from the OS of stream A into some dimensions of the OS of stream B.

The IS of the combined stream represents the join of the OS of two fact streams with complementary dimensions and measures.

The IS of the combined stream represents the union of complementary facts with the same multidimensional structure (i.e., equivalent data schemas).

The system will only accept workflows with a coherent combination of data streams according to the previous interpretations (see Figure 4).

3.3. Temporal Consistency

When designing workflows, temporal consistency must be checked in the combined data streams. Temporal consistency is associated to the speed of the streams. For example, a stream that is injecting dimension members/attributes to another fact stream must be slower than it, otherwise its aggregation power degrades. Performing joint and union operations on fact streams can also affect the temporal consistency: the speed of the joined stream cannot be faster than those of the input streams. Additionally, to make sense, time windows must be always longer or equal than the time slider. Finally, as the workflow is intended to get summarized facts, aggregations should be performed on increasingly coarser granularities.

It is worth mentioning that the formal representation of all these restrictions is beyond the scope of this paper. The natural way to make explicit an actionable representation of the presented models is with the OWL language [43], as most data in the streams are expressed as RDF graphs.

4. An Example Use Case

In this section we present a real-scenario use case to show how the previous architecture allow us to express the workflow of a complex streaming analysis problem. The problem we want to model consists of getting a proper insight from the tweets written by relevant actors within the automotive sector by considering current car brands and models. The chosen application scenario is related to companies whose business depends on the Car Industry. Specifically, business like rent-a-car should continuously monitor the user preferences and claims in order to renew the fleet of cars they should offer. With the proposed solutions, a Business Analyst can propose user profiles and specific metrics, which can be quickly deployed and monitored in the stream infrastructure. Thus, analysts can dramatically reduce the time spent in extracting and shaping social data, consuming directly the facts produced by the designed workflows. Figure 5 shows the proposed workflow for this problem, which is described in turn.

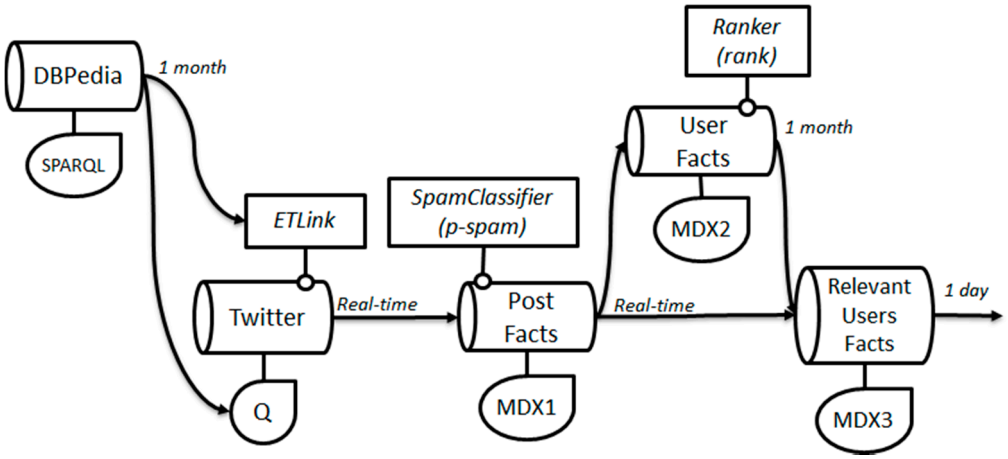


Figure 5. Example workflow for the use case scenario.

Firstly, we define a stream for identifying new car brands and models, which uses Live DBPedia via its SPARQL endpoint. The query is parametrized by using as reference the date of a month ago (parameter one_month_ago), which is set when fetching the query to the endpoint:

```
select ?car,?date,?brand where {
   ?car dbo:manufacturer ?brand.
   ?car dbo:productionStartYear ?date.
FILTER (?date > $one_month_ago$^^xsd:dateTime)
}
```

This stream feeds both the ETLink process that semantically enrich tweet data, and the Twitter query track, which includes the names of the car brands and models to be followed in Twitter. Whenever a new car model appears in DBPedia, the ETLink and Twitter query are accordingly updated. In this example, ETLink basically consists of a dictionary that maps text chunks to entities in a reference ontology [3]. An interesting indicator could measure the delay between the updates in DBPedia and the mentions in the tweets of car models.

Once facts are generated applying the ETLink process, these are processed to get incrementally the desired analytical data. The first fact stream consists of a PRE-process to automatically assign a probability to each fact to be a spam. Basically, this process consists of a classifier previously trained with recent historical data. It adds to each fact a new measure for reflecting its probability as spam. The MDX1 query then select facts with low spam probability. The resulting facts feed two streams with different processing speeds. In Figure 5 we show stream speeds at the output of each stream. In this example, time windows coincide with the time sliders.

The User fact stream summarizes a group of metrics concerning the user (e.g., followers, published tweets on domain, total published tweets, etc.) at month granularity to get a ranking of users. This process is reflected in the Ranker POST-process, which calculates the relevance of the user according to their aggregated metrics. Finally, another fact stream joins the incoming post facts with the user fact streams to select the non-spam facts of the relevant users (this join will be expressed in the MDX3 query). This is an example of attribute injection for one of the dimensions of the post facts.

The resulting stream workflow generates summarized data at day level, which is stored in the Serving Layer of the architecture to be consumed and visualized by means of the proper analytical tools.

Figure 6 shows the schema of the output of the Twitter stream after the ETLink is applied. We can see a constellation of two fact types, which can be joined by the User dimension. The multidimensional query MDX1 outputs the schema shown in Figure 7, where the User Fact dimensions become attributes

of the User dimension of Post Fact. These facts are again transformed at the User Fact stream, where the multidimensional query MDX2 shapes facts as follows: the number of the post facts associated to each user within the time window are summarized into a measure of User Fact (on_domain). Figure 7 also shows the derived measures and attributes from POST/PRE-processes with dotted lines. The final output multidimensional schema, which is fetched to the serving layer, is shown in Figure 8.

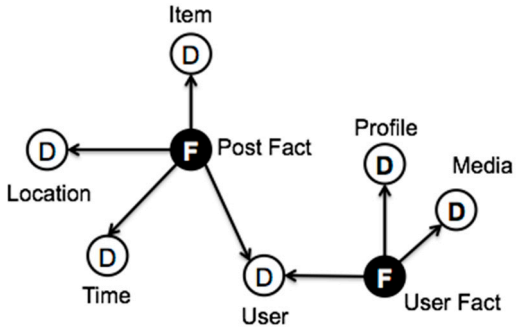


Figure 6. Graphs obtained after applying ETLink.

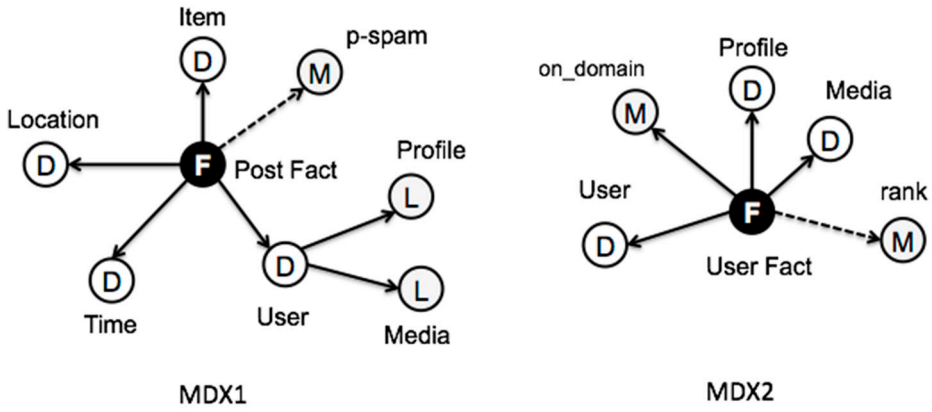


Figure 7. Output schemas after applying MDX1 and MDX2.

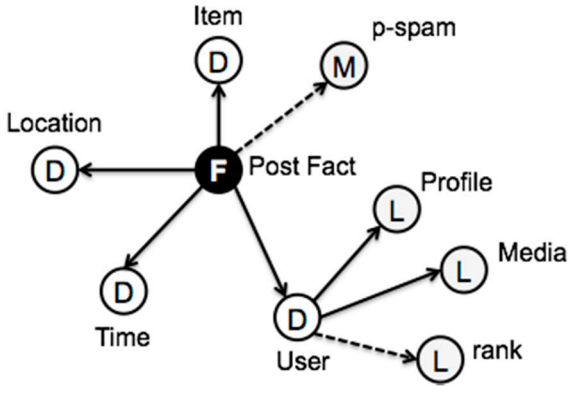


Figure 8. Output Schema after applying MDX3.

5. Prototype Implementation

We have implemented a small prototype with Python to show some analyses performed using this approach. The implementation of the source streams is quite straightforward with the available libraries for Twitter. We have also adopted libraries from SLOD-BI to implement the ETLink process for generating the tweet facts as shown in Figure 5. It is worth mentioning that ETLink processes are

very efficient since they rely on automatic semantic annotation methods that map text chunks and data to ontology entities. Moreover, ETLink processes can be easily parallelized so that they can be adapted to the speed of the incoming posts stream.

In the current prototype, each stream is implemented as a web service, whose data are pulled by the consumer streams. These web services maintain a cursor for each consumer stream. After all data in the current time window has been served to the consumer streams, cursors are closed and once data has been computed for a new time window, cursors are opened back for further processing. These stream services deal with JSON data and fact streams serve and consume data JSON-LD format [44]. In this way, for future implementations, it will be possible to use NoSQL databases for supporting large stream buffers. We also plan to automatically execute these workflows within a fully scalable Big Data framework such as Kafka and Spark.

Regarding the Spam Classifier, it has been trained and tested using the Anaconda framework (Pandas and ScikitLearn packages) [45], which is also implemented in Python so classifiers can be easily integrated into the stream services. Following the proposed architecture, we maintain a long-term stage store of one year from where obtain the necessary samples for training the Spam Classifier with a Linear SVM. After applying the Spam Classifier, the number of facts is reduced around 40%.

The Ranker process shown in Figure 5 has been implemented with a simple formula that is applied after user facts have been aggregated. This measure corresponds to the ordering of the on_domain and total number of tweets per user.

We have simulated the streaming workflow with a series of historical tweets about car models tracked during the year 2015. Figure 9 shows the results generated by the User Fact streams during that year (only the top 15 are considered). As it can be noticed, the users ranking is different at each month, and therefore it is necessary to maintain a relatively stable list of relevant users to perform interesting analysis with post facts. If it is not possible to obtain a stable user list then "user" is not a valid dimension for analysis.

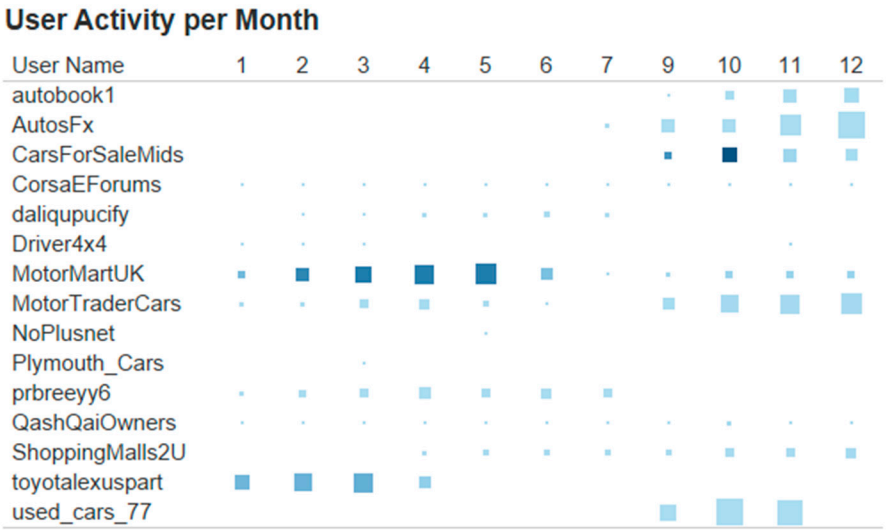


Figure 9. Relevant Users per activity. Size is proportional to the number of on domain tweets and color indicates the total number of posted tweets.

Figure 10 shows the number of on-domain tweets per brand generated by the initial ETLink process aggregated at the week level. In the figure we can clearly identify two main brands dominating the published tweets in this domain (i.e., Toyota, Aichi, Japan, and Ford, Dearborn, MI, USA). It must be said that the original post facts are associated to the names of car models, which are grouped by brand for visualization.

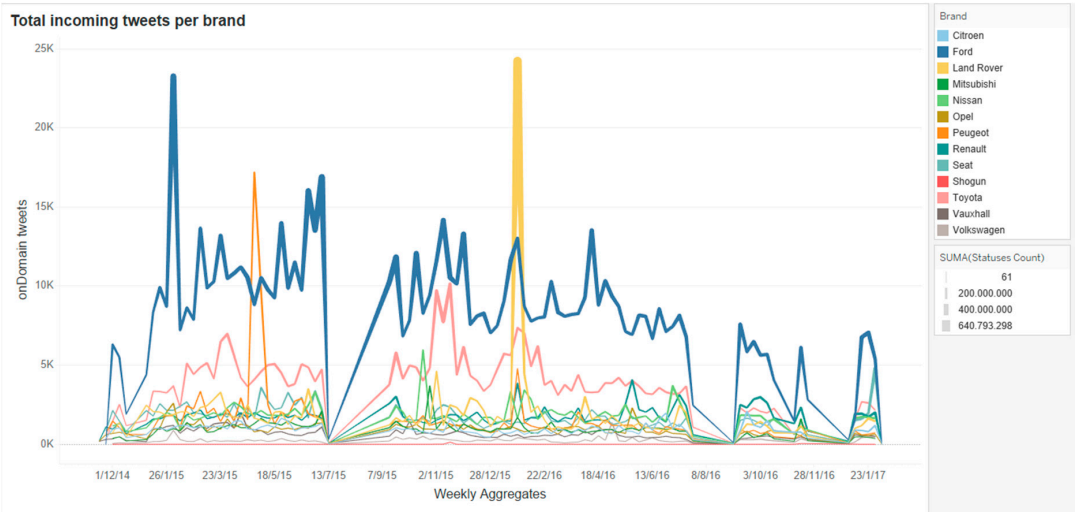


Figure 10. Total incoming post facts for the example use case.

Figure 11 shows the final output after selecting the top 15 relevant users shown in Figure 9, and after removing spam post facts. In this figure can see that the main relevant users focus on different brands than in the incoming fact stream of Figure 10. For example, Toyota brand paid a great attention during the first half of the year, whereas other brands competed with it during the second half of the year.

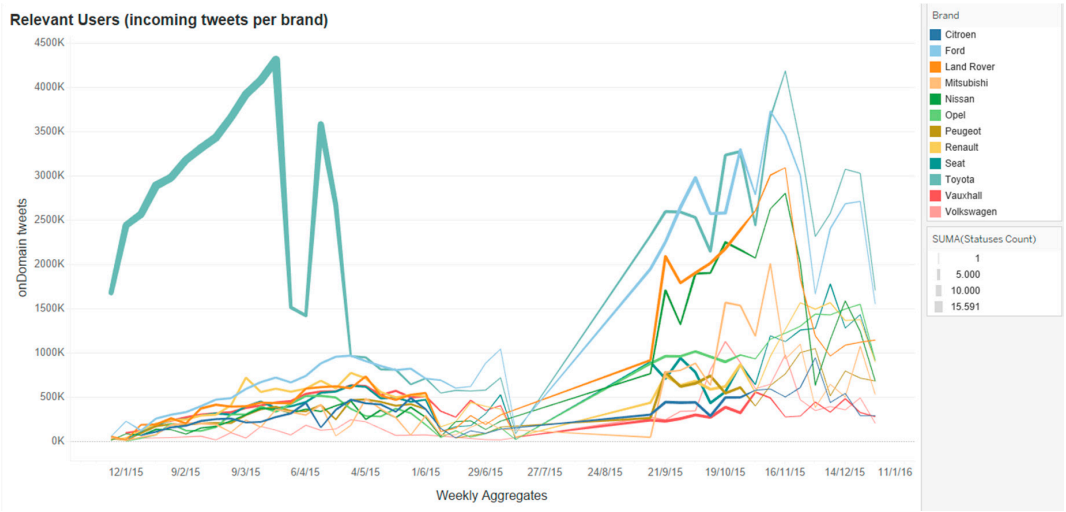


Figure 11. Post facts per brand for top 15 relevant users.

6. Conclusions

In this work we have analyzed a comprehensive set of methods related to Social Business Intelligence (SBI). We conclude that streaming is the natural way to process social data, and it implies new challenges for analytical tasks. The main challenge is the dynamicity of all the elements implied in the analyses, starting from the data sources and ending up with the analytical indicators. Another challenge is the need for intelligent processing in most of the analysis tasks associated to SBI, like sentiment analysis, spam detection, and so on. Indeed, we cannot think of a SBI task without including both analysis types: multidimensional and predictive.

We propose a new architecture that aims at covering all these requirements as well as at integrating Data Science and Data Analysis tasks in the same working area. We adopt a Kappa-like streaming architecture to cover the requirements of both kinds of actors. The architecture relies on both Linked Data and multidimensional modeling. The former easies the semantic data enrichment whereas the latter shapes them for analysis purposes. The adoption of semantics also facilitates the validation and follow-up of the developed workflows.

Future work has two main directions. One is to achieve a full implementation of the architecture in Python and its integration with a highly scalable streaming framework like Kafka. Another direction is to get a complete theoretical model for the stream workflows relying on Description Logics. The goal of this model is to keep an up-to-date and consistent logic representation of the workflow, which can be used to validate stream re-use and composition as well as the automatic execution of workflows.

Author Contributions: Conceptualization, R.B., I.L.-C. and M.J.A.; Methodology, R.B., I.L.-C. and M.J.A.; Software, R.B.; Validation, R.B. and I.L.-C.; Formal Analysis, R.B., I.L.-C. and M.J.A.; Investigation, R.B., I.L.-C. and M.J.A.; Resources, R.B. and I.L.-C.; Data Curation, R.B. and I.L.-C.; Writing-Original Draft Preparation, I.L.-C. writes the introduction and Section 2, R.B. writes Sections 3–5 and conclusions; Writing-Review & Editing, R.B., I.L.-C. and M.J.A.; Visualization, R.B. and I.L.-C.; Supervision, R.B.; Project Administration, R.B.; Funding Acquisition, the members of the project TIN2017-88805-R are R.B., I.L.-C. and M.J.A., I.L.-C. is the pre-doctoral student benefiting from the grant "PREDOC/2017/28".

Funding: This research was funded by the Spanish Ministry of Industry and Commerce grant number TIN2017-88805-R and by the pre-doctoral grant of the Universitat Jaume I with reference PREDOC/2017/28.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Inmon, W. *Building the Data Warehouse*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2005.
2. Kreps, J. Questioning the Lambda Architecture 2014. Available online: https://www.oreilly.com/ideas/questioning-the-lambda-architecture (accessed on 11 June 2018).
3. Berlanga, R.; García-Moya, L.; Nebot, V.; Aramburu, M.; Sanz, I.; Llidó, D. SLOD-BI: An Open Data Infrastructure for Enabling Social Business Intelligence. *Int. J. Data Warehous. Data Min.* **2015**, *11*, 1–28. [CrossRef]
4. Liu, X.; Tang, K.; Hancock, J.; Han, J.; Song, M.; Xu, R.; Pokorny, B. A Text Cube Approach to Human, Social, Cultural Behavior in the Twitter Stream. In Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, Washington, DC, USA, 2–5 April 2013.
5. Rosenthal, S.; Farra, N.; Nakov, P. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017.
6. Montejo-Ráez, A.; Martínez-Cámara, E.; Martín-Valdivia, M.; Ureña-López, L.A. Ranked Wordnet graph for sentiment polarity classification in Twitter. *Comp. Speech Lang.* **2014**, *28*, 93–107. [CrossRef]
7. Volkova, S.; Bachrach, Y.; Armstrong, M.; Sharma, V. Inferring Latent User Properties from Texts Published in Social Media. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; pp. 4296–4297.
8. Pennacchiotti, M.; Popescu, A.-M. A Machine Learning Approach to Twitter User Classification. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Catalonia, Spain, 17–21 July 2011; pp. 281–288.
9. Colleoni, E.; Rozza, A.; Arvidsson, A. Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *J. Commun.* **2014**, *64*, 317–332. [CrossRef]
10. Kapanipathi, P.; Jain, P.; Venkataramani, A.C. User interests identification on twitter using a hierarchical knowledge base. In Proceedings of the 11th European Semantic Web Conference ESWC 2017, Portorož, Slovenia, 28 May–1 June 2017.
11. Miller, Z.; Dickinson, B.; Deitrick, W.; Hu, W.; Wang, A. Twitter spammer detection using data stream clustering. *Inf. Sci.* **2014**, *260*, 64–73. [CrossRef]
12. Varol, O.; Ferrara, E.; Davis, C.; Menczer, F.; Flammini, A. Online Human-Bot Interactions: Detection, Estimation, and Characterization. Available online: https://arxiv.org/abs/1703.03107 (accessed on 11 June 2018).

13. Subrahmanian, V.; Azaria, A.; Durst, S.; Kagan, V.; Galstyan, A.; Lerman, K.; Zhu, L.; Ferrara, E.; Flammini, A.; Menczer, F. The DARPA Twitter Bot Challenge. *Computer* **2016**, *49*, 38–46. [CrossRef]
14. Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; Flammini, A. The Rise of Social Bots. *Commun. ACM* **2016**, *59*, 96–104. [CrossRef]
15. Li, H.; Mukherjee, A.; Liu, B.; Kornfield, R.; Emery, S. Detecting Campaign Promoters on Twitter using Markov Random Fields. In Proceedings of the IEEE International Conference on Data Mining, Shenzhen, China, 14–17 December 2014.
16. Roelens, I.; Baecke, P.; Benoit, D. Identifying influencers in a social network: The value of real referral data. *Decis. Support Syst.* **2016**, *91*, 25–36. [CrossRef]
17. Xie, W.; Zhu, F.; Jiang, J.; Lim, E.-P.; Wang, K. TopicSketch: Real-Time Bursty Topic Detection from Twitter. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 2216–2229. [CrossRef]
18. Feng, W.; Zhang, C.; Zhang, W.; Han, J.; Wang, J.; Aggarwal, C.; Huang, J. STREAMCUBE: Hierarchical spatio-temporal hashtag clustering for event exploration over the Twitter stream. In Proceedings of the 2015 IEEE 31st International Conference on Data Engineering, Seoul, Korea, 13–17 April 2015.
19. Zhang, C.; Zhou, G.; Yuan, Q.; Zhuang, H.; Zheng, Y.; Kaplan, L.; Wang, S.; Han, J. GeoBurst: Real-Time Local Event Detection in Geo-Tagged Tweet Streams. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, Pisa, Italy, 17–21 July 2016.
20. Zhou, X.; Chen, L. Event detection over twitter social media streams. *VLDB J.* **2014**, *23*, 381–400. [CrossRef]
21. Atefeh, F.; Khreich, W. A Survey of Techniques for Event Detection in Twitter. *Comput. Intell.* **2013**, *31*, 132–164. [CrossRef]
22. Zubiaga, A.; Spina, D.; Martínez, R.; Fresno, V. Real-time classification of Twitter trends. *J. Assoc. Inf. Sci. Technol.* **2015**, *66*, 462–473. [CrossRef]
23. Cao, G.; Wang, S.; Hwang, M.; Padmanabhan, A.; Zhang, Z.; Soltani, K. A scalable framework for spatiotemporal analysis of location-based social media data. *Comput. Environ. Urban Syst.* **2015**, *51*, 70–82. [CrossRef]
24. Smith, M.A. NodeXL: Simple Network Analysis for Social Media. In *Encyclopedia of Social Network Analysis and Mining*; Springer: New York, NY, USA, 2014.
25. Barbieri, D.; Braga, D.; Ceri, S.; Della Valle, E.; Huang, Y.; Tresp, V.; Rettinger, A.; Wermser, H. Deductive and Inductive Stream Reasoning for Semantic Social Media Analytics. *IEEE Intell. Syst.* **2010**, *25*, 32–41. [CrossRef]
26. Smith, M.A.; Shneiderman, B.; Milic-Frayling, N.; Mendes Rodrigues, E.; Barash, V.; Dunne, C.; Capone, T.; Perer, A.; Gleave, E. Analyzing (Social Media) Networks with NodeXL. In Proceedings of the Fourth International Conference on Communities and Technologies, New York, NY, USA, 25–27 June 2009; pp. 255–264.
27. Berlanga, R.; Aramburu, M.; Llidó, D.; García-Moya, L. Towards a Semantic Data Infrastructure for Social Business Intelligence. In *New Trends in Databases and Information Systems*; Springer: Cham, Switzerland, 2014.
28. Nebot, V.; Berlanga, R. Statistically-driven generation of multidimensional analytical schemas from linked data. *Knowl.-Based Syst.* **2016**, *110*, 15–29. [CrossRef]
29. Francia, M.; Gallinucci, E.; Golfarelli, M.; Rizzi, S. Social Business Intelligence in Action. In Proceedings of the Advanced Information Systems Engineering: 28th International Conference CAiSE, Ljubljana, Slovenia, 13–17 June 2016; pp. 33–48.
30. Scholl, S.R.H. Discovering OLAP dimensions in semi-structured data. *Inf. Syst.* **2014**, *44*, 120–133.
31. Mauri, A.; Calbimonte, J.; Dell'Aglio, D.; Balduini, M.; Brambilla, M.; Della Valle, E. TripleWave: Spreading RDF Streams on the Web. In Proceedings of the Semantic Web—ISWC 2016. ISWC 2016, Kobe, Japan, 17–21 October 2016.
32. Balduini, M.; Della Valle, E.; Dell'Aglio, D.; Tsytsarau, M.; Palpanas, T.; Confalonieri, C. *Social Listening of City Scale Events Using the Streaming Linked Data Framework*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 1–16.
33. Zeng, D.; Chen, H.; Lusch, R. Social Media Analytics and Intelligence. *IEEE Intell. Syst.* **2010**, *25*, 13–16. [CrossRef]
34. Nadal, S.; Herrero, V.; Romero, O.; Abelló, A.; Franch, X.; Vansummeren, S.; Valerio, D. A software reference architecture for semantic-aware Big Data systems. *Inf. Softw. Technol.* **2017**, *90*, 75–92. [CrossRef]
35. Stonebraker, M.; Çetintemel, U.; Zdonik, S. The 8 Requirements of Real-Time Stream Processing. *SIGMOD Rec.* **2005**, *34*, 42–47. [CrossRef]

36. Marz, N.; Warren, J. *Big Data: Principles and Best Practices of Scalable Realtime Data Systems*, 1st ed.; Manning Publications Co.: Greenwich, CT, USA, 2015.
37. Javed, M.H.; Lu, X.; Panda, D.K. Characterization of Big Data Stream Processing Pipeline: A Case Study using Flink, Kafka. In Proceedings of the Fourth IEEE/ACM International Conference on Big Data Computing, Applications, Technologies, New York, NY, USA, 5–8 December 2017; pp. 1–10.
38. Hebeler, J.; Fisher, M.; Blace, R.; Perez-Lopez, A. *Semantic Web Programming*; John Wiley & Sons: Hoboken, NJ, USA, 2009.
39. DBPedia Live. Available online: https://wiki.dbpedia.org/online-access/DBpediaLive (accessed on 20 June 2018).
40. BabelNet Live. Available online: http://live.babelnet.org/ (accessed on 20 June 2018).
41. Romero, O.; Abelló, A. A framework for multidimensional design of data warehouses from ontologies. *Data Knowl. Eng.* **2010**, *69*, 1138–1157. [CrossRef]
42. Barbieri, D.F.; Braga, D.; Ceri, S.; Della Valle, E.; Grossniklaus, M. Querying RDF streams with C-SPARQL. *SIGMOD Rec.* **2010**, *39*, 20–26. [CrossRef]
43. OWL Language. Available online: https://www.w3.org/OWL/ (accessed on 20 June 2018).
44. JSON-LD. Available online: https://json-ld.org/ (accessed on 20 June 2018).
45. Anaconda. Available online: https://anaconda.org/ (accessed on 20 June 2018).

Chapter 4

Multidimensional Author Profiling for Social Business Intelligence

Publication

Lanza Cruz, Indira Lázara, Berlanga, Rafael and Aramburu, María José. Multidimensional Author Profiling for Social Business Intelligence. Inf Syst Front (2023). Springer. https://doi.org/10.1007/s10796-023-10370-0. (Q1)

Information Systems Frontiers
https://doi.org/10.1007/s10796-023-10370-0

Multidimensional Author Profiling for Social Business Intelligence

Indira Lanza-Cruz[1] · **Rafael Berlanga**[1] · **María José Aramburu**[2]

Accepted: 3 January 2023

Abstract
This paper presents a novel author profiling method specially aimed at classifying social network users into the multidimensional perspectives for social business intelligence (SBI) applications. In this scenario, being the user profiles defined on demand for each particular SBI application, we cannot assume the existence of labelled datasets for training purposes. Thus, we propose an unsupervised method to obtain the required labelled datasets for training the profile classifiers. Contrary to other author profiling approaches in the literature, we only make use of the users' descriptions, which are usually part of the metadata posts. We exhaustively evaluated the proposed method under four different tasks for multidimensional author profiling along with state-of-the-art text classifiers. We achieved performances around 88% and 98% of F1 score for a gold standard and a silver standard datasets respectively. Additionally, we compare our results to other supervised approaches previously proposed for two of our tasks, getting very close performances despite using an unsupervised method. To the best of our knowledge, this is the first method designed to label user profiles in an unsupervised way for training profile classifiers with a similar performance to fully supervised ones.

Keywords Social media · Author profiling · Business intelligence · Natural language processing · Machine learning

1 Introduction

Social business intelligence (SBI) is the field that combines corporate data with user-generated contents with the aim of improving decision making within the companies (Berlanga et al., 2015; Gallinucci et al., 2015). Social business intelligence aims at analysing from different perspectives the texts that users generate as well as their reactions and interactions. Thus, social networks become a rich source of immediate information that can provide useful insights for businesses. Unfortunately, the main challenge of SBI lies in the effective management of the social network's contents, which are very noisy, unstructured and dynamic by nature. Even when retrieving user posts with specific queries, the number of non-relevant and off-domain posts is too large to extract reliable and useful information.

In this paper, we apply author profiling (AP) in order to characterize both the contents generators and the audience that is interacting with these contents. This information is very useful to increase the quality of social media data for many different purposes (Aramburu et al., 2021). Our main aim is to build dynamic analysis dimensions that allow analysts to focus on particular user profiles according to their current goals. For example, an analysis whose target is to study the promotions in social media of certain products will mainly focus on authors with a professional profile. Profiles are also multidimensional by nature, that is, an author can be characterized from different perspectives like demographics, business roles, domains of interest, and so on.

In our previous work (Aramburu et al., 2021; Lanza-Cruz et al., 2018), we performed the classification of profiles by means of manually selected keywords. Although this method provided us the necessary information for quality assessment and global statistics, it is clearly deficient for multidimensional analysis because of its poor coverage. In this paper, we introduce a formal methodology that allows building the necessary profile classifiers from the analysts' specifications.

✉ Indira Lanza-Cruz
indiralanza@yahoo.es

[1] Department de Llenguatges i Sistemes Informàtics, Universitat Jaume I, Avda. Vicent Sos Baynat s/n, Castellón de la Plana, Spain

[2] Department d'Enginyeria i Ciència dels Computadors, Universitat Jaume I, Avda. Vicent Sos Baynat s/n, Castellón de la Plana, Spain

Published online: 06 February 2023

Previous research in AP for social networks has been mainly focused on inferring from the users' posts aspects like age, gender and language variety (Daelemans, et al., 2019; Pardo et al., 2016; Potthast, et al., 2019; Weren et al., 2014. Other studies aimed at specific user profiles such as campaign promoters, bots, influencers, political position and polarity (Amigó et al., 2014; Li et al., 2014; Nebot et al., 2018; Pennacchiotti & Popescu, 2011), among others. Most of these approaches rely on machine learning techniques, usually involving strong feature engineering efforts and the preparation of large training data (McCorriston et al., 2015; Mishra et al., 2018; Wood-Doughty et al., 2018). In a dynamic scenario, these approaches are not feasible because for any new profile we want to include in a SBI analysis, we would require the definition of a new training dataset. As far as we know, there is no approach able to deal with multiple and dynamic dimensions for user profiles as proposed in this paper.

In our approach, we start with a formal description of the multidimensional model associated with the user profiles of interest. Then, from this model, we automatically generate the datasets for training the corresponding author profile classifiers. In this way, analysts only intervene during the specification of the multidimensional model. Furthermore, analysts can evolve the multidimensional user profiles on demand by redefining the formal descriptions, and the system will automatically update the corresponding classifiers with the newly generated samples.

1.1 Research Objectives

The main objective of our research is to propose a novel approach to quickly constructing automatic profile classifiers suited to user-defined multidimensional SBI analysis perspectives.

Our main hypothesis is that we can automatically generate high-quality labelled datasets from unlabelled ones by applying two kinds of knowledge-based techniques, namely: word embeddings and ontologies. Thus, we can directly use the generated dataset to train the intended profile classifiers for the user-defined classes.

The specific objectives of our approach are as follows:

- Define a formal description of the multidimensional model associated with the users' profile classes.
- Design a method for the extraction of semantic key bigrams from the analyst specifications so that a labelled dataset of user profiles can be generated.
- Train and evaluate different user profile classifiers with the generated datasets.

1.2 Organization of the Paper

The rest of the paper is organized as follows. Section 2 discusses the related work and Section 3 proposes a multidimensional analysis model to classify user profiles. Section 4 presents a novel method to construct tagged datasets for author profiling with minimal human supervision. Section 5 describes the material used to develop our proposal, and Section 6 the methods used and the experimental settings. Specifically, the proposed classification tasks and the evaluation measures of the models are explained. Section 7 describes the model's validation approaches and discusses the results. Finally, Section 8 presents the conclusions and future work.

2 Related Work

The problem of AP has been addressed in several scientific branches, especially in the fields of linguistics and natural language processing (NLP) with approaches based on the analysis of textual features extracted from documents. Classical AP approaches have proven to be effective on formally written texts, such as books or the written press. They combine textual attributes ranging from lexical features (e.g., content words and function words) to syntactical features (e.g., POS-based features). Author profiling approaches for social networks show that the most useful attributes are statistical combinations of textual content (e.g., discriminative words) and stylistic features (Schler et al., 2006; López-Monroy et al., 2015). Previous text classification methods combine textual information with other features such as the post metrics (e.g., followers, likes, etc.), user interaction patterns (McCorriston et al., 2015; Kim et al., 2017), as well as sentiment and temporal features for the detection of bots and promoters (Daelemans, et al., 2019; Li et al., 2014; Ouni et al., 2021). Other paper proposes an hybrid approach for identifying the spam profiles by combining social media analytics and bio inspired computing (Aswani et al., 2018). A different approach is used to categorize hacker users in online communities where users are grouped according to their posting patterns (Zhang et al., 2015).

Author profiling in social networks has been mainly addressed for the identification of demographic attributes such as age, gender and geolocation (Daelemans, et al., 2019; López-Santillán et al., 2020; Ouni et al., 2021; Rangel et al., 2021; Schlicht & Magnossão de Paula, 2021; Young et al., 2018). Other complex attributes such as personality traits and influence degree have been also treated

extensively in the literature (Cervero et al., 2021; Kumar et al., 2018; Nebot et al., 2018; Rodríguez-Vidal et al., 2021; Schler et al., 2006; Weren et al., 2014). Its relevance and applicability in real-world scenarios can be seen in the different editions of the PAN International Competition on AP (Daelemans, et al., 2019; Pardo et al., 2016; Potthast et al., 2019).

For three years, the CLEF initiative RepLab has held a competitive evaluation of tasks in online reputation management. Among many relevant tasks, they developed solutions for Twitter author profiling, opinion makers identification, and reputational dimension classification (Carrillo-de-Albornoz et al., 2019). In the RepLab 2014 edition (Amigó et al., 2014), one of the objectives was author profiling in the automotive and banking domains. This initiative gives us a complementary view to PAN competition, as AP is more focused to the SBI perspective. In their solutions, researchers made use of a combination of features such as quantitative profile metadata, stylistic and behavioural features. Unfortunately, the obtained results did not exceed 50% in accuracy and MAAC evaluation measures, very close to the baselines. Despite the results, we consider this work a first approach to author profiling in Twitter, which served as the basis for further research (Potthast et al., 2019).

As previously mentioned, most approaches in AP have relied on machine learning with a strong feature engineering effort. Instead, recent techniques based on neural networks for text processing have allowed researchers to define easily new text classifiers as end-to-end solutions. However, only a few works involve the use of word/document embeddings in AP tasks. Markov et al. (2017) applied document embeddings to improve the classification performance in the PAN 2016 competition. López-Santillán et al. (2020) proposed a method to generate documents embeddings by means of evolving mathematical equations that integrate words frequency statistics, namely term frequency, TF-IDF, Information Gain and a new feature called relevance topic value. They employed Genetic Programming to weight word embeddings produced with different methods such as word2vec (Mikolov et al., 2013), fastText (Bojanowski et al., 2017) and BERT (Devlin et al., 2019). Then, they created a document embedding with their weighted averaging. They evaluated their proposed method over PAN's datasets (from 2013 to 2018) to predict personal attributed of authors. Schlicht and Magnossão de Paula (2021) proposed a framework to identify hate speech spreaders by applying several variants of BERT sentence transformers over the users' tweets. They also apply an attention mechanism to select important tweets for learning user profiles. For the classification task, they use the results of the PAN Profiling Hate Speech Spreader Task 2021 (Rangel et al., 2021) as a base line, improving them.

Another interesting perspective for SBI is the classification of profiles into individuals and organizations. To the best of our knowledge, three proposals have addressed this task in the literature. McCorriston et al. (2015) presented a study of organization demographics and behavior on Twitter, called Humanizr. They proposed a text classifier to distinguish between personal and organizational users. The training dataset was manually tagged through a crowdsourcing platform. As in PAN and RepLab, each tagged user is associated to a fixed number of posted tweets (in this case 200). Wood-Doughty et al. (2018) proposed Demographer, a tool for demographic inference of Twitter users. They emphasized the need of approaches that optimize computing resources in terms of time, number of data elements, and API accesses. Thus, they applied a minimum number of features in their predictive tasks, instead of a fixed number of posted tweets. For classifying person vs. organization users, they make use of n-grams and neural models trained with user names and profile features. They also proposed a series of ad-hoc methods to build a large labelled dataset in a semi-supervised way. For example, they extract individual accounts from Twitter lists organized by topic and use keywords such as "business" or "companies" to identify organization accounts. The dataset is manually verified by taking random samples. Finally, Wang et al. (2019) presented the method M3, aimed to identify gender, age and person-vs-organization users. M3 is a deep learning system that infers demographic attributes from social media profiles. This system only uses the user profile data, namely: profile image, username, screen name and biography. The model architecture comprises two separate models, a DenseNet for images, and a character-based neural network model for texts. This is a fully supervised approach and therefore requires a great effort for preparing the training datasets.

Table 1 shows as a summary the attributes most frequently used in the literature for the task of AP in social networks. The table separates the features extracted from the Post and User objects to train the predictive models. In general, results show that approaches relying on users' descriptions allow generating larger training datasets and improve notably the classification results.

As main conclusion, existing AP approaches are time-consuming and demanding in terms of human resources, which is unfeasible for scenarios where classes must be dynamically redefined on demand like in SBI. In this work, we advocate the use of unsupervised methods for the construction of the training datasets and its combination with effective supervised methods for the continuous classification of profiles.

Table 1 Features the literature applied for social network author profiling

	Post				User			
	Text	Style	Polarity	Semantics	Metrics	Name	Image	Bio
Wang et al. (2019)						√	√	√
McCorriston et al. (2015)	√	√			√			
Wood-Doughty et al. (2018)	√	√			√	√		
López-Monroy et al. (2015)	√	√	√					
Pardo et al. (2016)	√	√						
Markov et al. (2017)				√				
Amigó et al. (2014)	√	√	√		√			
Nebot et al. (2018)	√			√				
Kim et al. (2017)	√			√	√		√	√
López-Santillán et al. (2020)	√			√				
Schlicht and Magnossão de Paula (2021)	√			√				√
(Zhang et al. (2015)	√	√						
(Aswani et al. (2018)	√	√	√	√	√			
(Kumar et al. (2018)	√			√				
Our approach								√

3 Knowledge-Centric Data Analysis

Social business intelligence mainly adopts the successful multidimensional data (MD) models defined for data warehouses (Kimball & Ross, 2013). MD models represent analytical data in the form of data cubes, which consist of dimensions for representing contexts, and measures for calculating indicators. Each position in a data cube is a fact and may contain several measures. Dimensions can be hierarchically organized into levels, which allow analysts to aggregate measures from different perspectives and detail levels. In (Berlanga et al., 2015) we proposed SLOD-BI, a multidimensional model for analysing social networks where two kinds of facts are regarded, namely: post facts and social facts. Social facts account for users and their interactions whereas post facts account for the user-generated contents. In this paper, we only focus on the social facts and therefore on the dimensions and metrics for users. Figure 1 shows different dimensions commonly associated with users as well as the new extended dimension for user profiling. This profile dimension consists of the necessary levels with which we can model users from complementary perspectives.

In the example of Fig. 1, we propose three levels, namely: Domain role, Group role and Business role. Domain role indicates whether the user's activities belong to the application domain or not. The Group role indicates whether the user represents an individual or a collective entity. Finally, the Business role indicates the kind of professional activities the user is involved in.

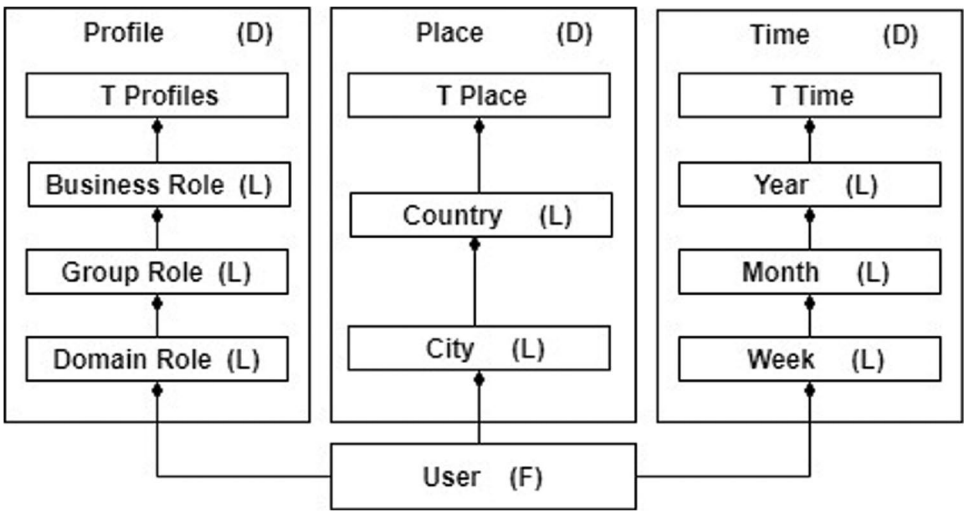


Fig. 1 Dimensions (D) and hierarchy levels (L) for representing user facts (F)

Each profile dimension level contains the different labels with which a user can be associated. For example, the Domain role contains two labels: on-domain and off-domain. Additionally, a dummy label is included for the unknown label, which is represented with "*" for simplicity.

Levels are ordered according to the analyst's requirements. In our example, we order levels according to their detail degree, namely: Domain role > Group role > Business role. In this way, we represent the specific labels of each level as a path following the specified order. For example, a journalist of the automotive domain is represented with the path 'journalist.individual.on-domain' whereas a cars magazine is represented as 'journalist.collective.on-domain'. From now on, we use the letter "i" for "individual", "c" for "·collective" and "a" for "on domain" to write concisely the paths of the multidimensional model.

When modelling the profile dimension, analysts must specify a description for each of the dimension level labels. These descriptions are used as the starting point to identify samples associated with the different profiles. Labels paths can be also described when they clearly denote an entity (e.g., a magazine as journalist.c.*).

Being part of SLOD-BI, all data is represented in a semantic framework. Thus, social facts are represented as ontologies. Dimensions and levels are represented as classes, which are properly related to represent the intended multidimensional model. In this way, this framework enables us to introduce further knowledge about the user profiles. For example, we can indicate that the class journalist is a subclass of the class professional. This knowledge can be very useful in case of label conflicts where the most general class can be taken as default.

Table 2 shows an example of users profile model. For our running example, we model six types of users for a Twitter data stream related to the automotive domain. Table 3 shows the mapping of the multidimensional hierarchy described in Table 2 with the class labels of the classifier model. We also include examples of user descriptions for these categories. The multidimensional hierarchy proposed to model the categories of users in Twitter allows us to represent in a simple way semantic relationships between users, organizations,

Table 2 Examples of user profile model dimensions for the automotive domain

Level	Value	Description / [*Seed unigrams with super-senses*]
Business role	Journalist	Profiles dedicated to the publication of news, such as journalists, newspapers, and magazines. / [*magazine: communication*] A journalist is a person who writes news to be broadcast. / [*journalist: person*]
	Professional	Professional company: It covers official companies, dealers, professional groups. / [*car: artefact, company: group*] Professional person: A professional is a member of a profession or any person who earns their living from a specified professional activity. Students, amateur, retirees, and former professionals are excluded from this classification. / [*professional: person*]
	Public Service	A service performed for the benefit of the community or its institutions. / [*police: group, services: group*]
	Lovers & fans	Person or groups of persons that use social networks as a means of learning, fun and entertainment. / [*lover: person, fan: person*]
Group role	Individual	A person. / [*me: person, I: person, my: person*]
	Collective	Any number of entities (members) considered as a unit. / [*we: group, our: group*]
Domain role	On domain	Indicates if the profile is related to the domain of analysis. / [*car: artefact*, …]
	Off domain	Domain different from the one analyzed.

Table 3 Example user descriptions of a multidimensional profile model. The text has been converted to lowercase

Label Path	Profile description example
professional.collective.on-domain	"business car specialists. lease, hire or buy, new or used cars and vans. supplying all types of vehicles. we offer business finance even if you have poor credit"
journalist.individual.*	"im an architect, designer, tv presenter, traveller and founder of the home educational charity mobie. all posts are my own."
journalist.collective.*	"official twitter page of the nation newspapers, nigeria's widest circulating newspaper. need to reach us?"…
journalist.individual.*	"reporter @kdvr & @channel2kwgn news in denver. travel junkie. dog & cat mom. patio beers."
publicService.collective.*	"official police scotland feed for ek, rutherglen & cambuslang. not for reporting crime. non-emergency calls dial 101 & 999 in an emergency. not monitored 24/7."
lovers&Fans.individual.on-Domain	"just your average teenage petrol head, trainee mechanic and overall motoring enthusiast"

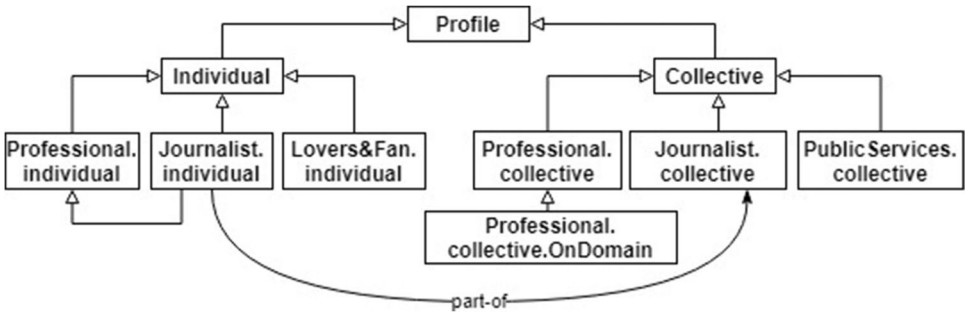


Fig. 2 Example multidimensional model expressed as an ontology

and topics. This type of analysis and hierarchical labelling can describe many types of relationships that analysts can query from different perspectives during the development of new applications.

We codify the multidimensional model with the ontology description language OWL-DL (W3C Recommendation OWL, 2004). To represent the semantic relationships between concepts we have reused some properties of the Simple Knowledge Organization System (SKOS) (W3C Semantic Web, 2012) resource. Figure 2 shows the relationships between classes, where the generalization-specialization relationships represent the broader and narrower properties respectively.

4 Learning Models for Author Profiling

Task The very goal of our work is to build a general, scalable and robust machine learning framework for large-scale classification of social network users according to a given multidimensional profile model as described in Section 3.

Solution Overview Figure 3 shows the proposed framework, which consists of three main components. The first component is an unsupervised method to generate tagged user profiles (silver standard dataset, SS) based on some analysis dimensions. The profiles labelling method makes use of semantic knowledge and word embeddings tools for the construction of a language model per defined class. It takes as input a stream of Twitter user profiles and a small set of tagged words according to classes and context (defined through WordNet (Miller, 1995) semantic groups). The second component is a Dataset Debugger aimed at cleaning and fining errors according to a small set of knowledge rules. The third component is a machine-learning algorithm that learns a classification model from the updated dataset (SS') which can be used to label a large dataset of user descriptions. The final step is the validation phase with a manually built gold standard dataset (GS).

4.1 The Method

The method we propose to generate labelled samples consists in iteratively identifying key bigrams for each target class of the multidimensional profile model (described in Section 3). The starting point is a set of seed unigrams (U_c^0) for each class. The seed unigrams are eventually selected by the domain experts, who first pick up unigrams from the profile class names and descriptions. Optionally, knowledge resources like Wikipedia.org, BabelNet.org and relatedwords.org can be used by experts to get further interesting unigrams. In this way, experts can face the cold start issue of the unsupervised phase. Table 2 shows some examples of seed unigrams which are mainly associated to the classes

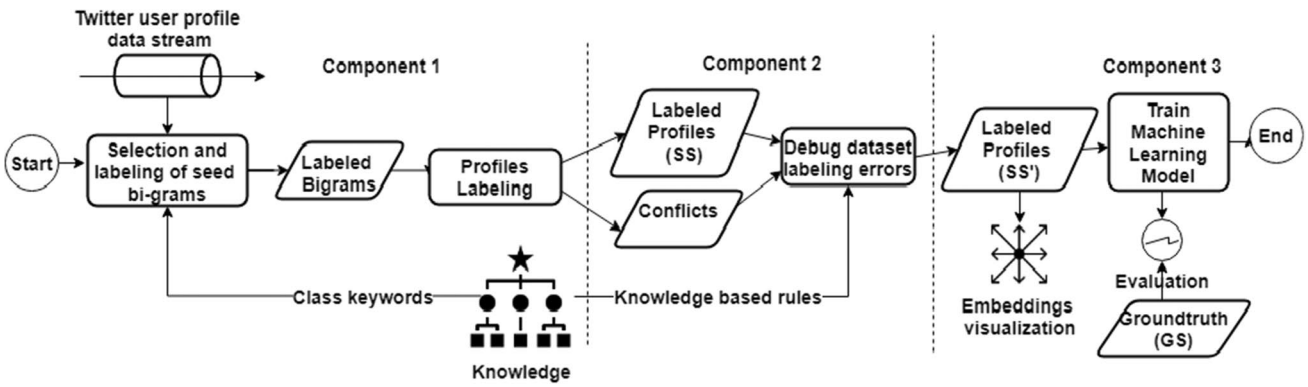


Fig. 3 Overview of the proposed workflow for the AP task

name and description. We also assume that there is a corpus $\mathbb{C}$ of unlabelled user descriptions, preferably related to the domain of analysis. Such a collection is easy to obtain from Twitter by fetching queries related to the domain of analysis. Let $\mathbb{B}$ be the set of all bigrams (u, v) with a skip of one word occurring in $\mathbb{C}$.

The goal of our method is to find a set of key bigrams $B_c \subseteq \mathbb{B}$ for each profile class c. The datasets B_c must be disjoint to each other in order to induce a partition in the unlabelled samples (clustering hypothesis) by means of representative and unambiguous bigrams. This assumption does not mean that a bigram in B_c will always be associated with class c. As mentioned in Section 4.2., conflict resolution can change the class of a sample that was initially assigned to a class c due to a bigram. It must be also pointed out that bigrams of different classes can share unigrams. These properties allow the method to reduce the bias of the output labelled dataset.

We assume that there exists a semantic similarity function between unigrams. Similarity between bigrams is a function that combines the similarity of their components. We also need a function to measure the similarity between a bigram and a set of unigrams.

Unigrams and bigrams are related through synonymy $(x \sim y)$ and entailment $(x \rightarrow y)$ relationships. Synonymy can be applied to pairs of either unigrams or bigrams, whereas entailment is applied between a unigram and a bigram. Some examples of these relationships are as follows:

$$car \backsim vehicle$$
$$(used, car) \backsim (used, vehicle)$$
$$car \rightarrow (used, vehicle)$$

In our approach, the synonymy relationship is weaker than those defined in lexical resources like WordNet. Synonymy mainly account for highly similar unigrams according to some provided semantic similarity function. The entailment relationship indicates whether a bigram can be semantically derived from a unigram or not.

For extracting the key bigrams of each class, we define the following procedure:

- For each class c:
  - Let U_c be the list U_c^0 of seed unigrams of class c along with all their synonymous unigrams.

    $$U_c = \left\{u | \exists v \in U_c^0 \, : \, u \sim v\right\}$$

  - For each unigram $u \in U_c$, we define the set of bigrams associated to u and related to c as follows $B_u^c = \left\{b | b \in \mathbb{B}, \exists u \in U_c \, : \, u \rightarrow b \wedge \nexists c', c' \neq c \, : \, sim\left(b, U_{c'}\right) > sim\left(b, U_c\right)\right\}$
- Finally, the set of key bigrams is defined as $B_c = \bigcup_{u \in U_c} B_u^c$

Table 4 Examples of bigrams associated to multi-level classes

Bigrams	Multi-level Class
(weekly, magazine)	journalist.collective.*
(journalist, editor)	journalist.individual.*
(lover, cars)	loversfans.individual.on-domain
(cars, services)	professional.collective.on-domain
(artistic, director)	professional.individual.*
(police, service)	publicservice.collective.*

Once key bigrams are determined for each class, we must ensure that the sets B_c are disjoint from each other. Thus, if a bigram occurs in more than one class, then we assign it to its most similar class. In case we cannot decide its most similar class, then we discard the bigram. Table 4 shows an example of bigrams associated to multi-level profile classes.

4.1.1 Semantic Synonymy

In this paper, we propose the use of word embeddings and WordNet super-senses (Ciaramita & Johnson, 2003) to define the similarity and synonymy of unigrams and bigrams. Similarity between unigrams is calculated with fastText by using the cosine of the two unigram-embedding vectors:

$$sim(x, y) = cosine\left(x_e, y_e\right)$$
$$u \sim v \;\; iff \;\; cosine\left(u_e, v_e\right) > \delta_K$$

Here x_e and y_e are the embedding vectors of x and y respectively. The threshold δ_K is set to the similarity of the K-most similar unigram for u. The value of K is a parameter that we will determine during model evaluations. We also train fastText to calculate the unigrams embedding vectors from the pool of profile descriptions ($\mathbb{C}$).

We define the entailment relationships in terms of the synonymy relationship along with the WordNet super-senses as follows:

$$u \rightarrow (v, w) \; iff \; (u \sim w \wedge ss(u) \cap ss(w) \neq \emptyset)$$
$$\vee (u \sim v \wedge ss(u) \cap ss(v) \neq \emptyset)$$

Where $ss(u)$ returns all super-senses associated to the unigram u. In this definition, the super-sense condition aims at constraining the semantics of the bigram head with respect to the unigram that entails it. As a further constraint, we can indicate which super-senses are allowed at each profile level in order to improve the quality of generated bigrams. For example, for the collective class, restricting the allowed super-senses to "organization" entails more precise bigrams.

Finally, the similarity between sets/sequences of unigrams is defined as the pair-wise average similarity of their components. Thus, the similarity between a bigram b and a unigram set U is the average of all pair-wise unigrams of the bigram and the set.

Table 5 Examples of conflicts between bigrams in the same user description

Bigram A	Class A	Bigram B	Class B
(reporter, covering)	journalist.i.*	(automotive, news)	journalist.c.a
(writer, editor)	journalist.i.*	(car, magazine)	journalist.c.a
(software, engineer)	professional.i.*	(cars, lover)	loversfans.i.a
(producer, reporter)	journalist.i.*	(auto, expert)	professional.i.a
(journalist, writer)	journalist.i.*	(expert, writer)	professional.i.*

conflict resolution. Table 5 shows several conflicting bigrams that jointly appear in users descriptions.

In the automotive domain, the most conflicting classes are 'journalist.i.*', and 'loversfans.i.*'. This is because many users describe their primary occupation along with their hobbies and interests, being journalism-related terms like reporter, editor, writer and blogger the most frequent ones in their descriptions. Additionally, we should consider classes related to the interest domain (e.g., car lovers) as complementary in presence of other classes. Next section explains how these implicit rules are applied to solve some conflicts.

Table 6 Rules based on the semantic relationship between concepts

Level	Specific Concept	Property	General Concept	Decision
Group role	Individual	PartOf	Collective	Individual
Domain role	Domain	PartOf	All	Domain
Business role	Journalist	Narrower	Professional	Journalist
	Professional	Narrower	PublicService	Professional
	PublicService	Narrower	LoversFans	PublicService

$$sim(b, U) = \frac{1}{2 \cdot |U|} \sum_{i=1}^{2} \sum_{u \in U} sim\left(b_i, u\right)$$

4.1.2 Data Labelling

Once we find the key bigrams B_c for each class c, we can assign labels to each sample $d \in \mathbb{C}$. The labeling method is as follows:

1. Let $labels(d)$ be all classes that have at least one key bigram in the sample d. Besides, this set must fulfill the following condition:

$$\forall c \in labels(d) : \nexists c', c' \notin labels(d) : sim\left(d, U_c\right) > sim(d, U_{c'})$$

2. We estimate the label distribution for each class $c \in labels(d)$, that is, $P\left(c|d, B_c\right)$.
3. We assign to the sample d the most likely labels for each profile level according to the previous distributions.

The condition in the first step ensures that each sample has its more likely class within the candidate classes derived from its bigrams.

If a sample contains more than one label in the same profile level, then we say that the sample contains a conflict. Some of these conflicts can be solved by applying the knowledge rules included in the multidimensional profile model (see Section 3). The defined rules must be designed to resolve frequent conflicts. Conflicts that cannot be resolved with the defined rules are excluded from training. In the next section, we show some examples of knowledge-based

Finally, all conflict-free labelled samples constitute the silver standard (SS) dataset, with which profile classifiers can be trained.

4.2 Detecting and Fixing Potential Labelling Conflicts

In order to increase the variety and complexity of the SS dataset, and considering that only conflict-free samples are regarded in the SS, we need to introduce some heuristics to solve frequent conflicts. More specifically, we make use of heuristics based on rules (Zanakis & Evans, 1981) to solve these frequent conflicts.

Given a conflict, a general rule of thumb is to assign the category of the most restrictive class, both at the class level and at the level of dimension granularity (see Table 6). The vocabulary used in the descriptions also gives clues about the specificity of class labels. Thus, vocabulary related to the domain will be much more frequent than that of some specific classes like journalist. Therefore, bigram frequencies provide us an easy way to decide which bigram is the most specific in a conflict.

We must point out that the quality of the result of the previous heuristic rules will depend on the quality of the bigrams associated with each class. On the other hand, the final decision can be modified depending on the perspective or dimension of the data we want to represent in our model. Low-frequency conflicts do not need to be analyzed, as they do not greatly determine the accuracy of the model. In this case, we should disregard samples with conflicts.

As an example of conflict resolution, the profile description "*senior news reporter, daily star newspaper. views own.*

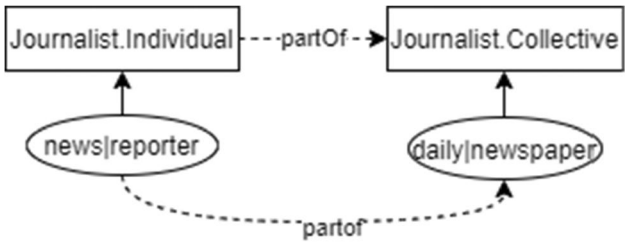


Fig. 4 Conflict resolution for individual vs. collective classes

Table 7 Total of unlabelled profiles obtained for each analysis domain

Cars	Tourism	Medicine
1,074,826	16,484,323	502,916

(email)." contains two conflicting bigrams: (news, reporter) and (daily, newspaper). The final decision is 'journalist.i.*' according to the rules shown in Fig. 4.

Finally, we also include a set of labelled unigrams that unambiguously imply a specific class or category. For example, frequent words like 'I' and 'me' imply the *individual* category. These unigrams allow us to detect some implicit conflicts that cannot be detected through bigrams. It is must be pointed out that only a few of these unigrams were included in the experiments to detect some frequent implicit conflicts.

5 Material and Statistics

To demonstrate the usefulness of the proposed method for AP and its applicability to different domains of analysis, we developed a series of experiments on three domains, namely: cars, tourism and medicine domains. For each domain we collected user profiles over a long-term stream of tweets, which has served as a basis of several analytical studies about SBI with very different proposals (Aramburu et al., 2020, 2021; Lanza-Cruz et al., 2018).

We generated each domain data stream with a series of keywords related with the corresponding domain. For example, for the automobile domain we used keywords representing different car models and brands. The data streams allowed capturing profiles comprising people who posted and interacted with the tweets. Table 7 shows the total number of unlabelled profiles obtained for each analysis domain. The dataset is preprocessed to replace all the links, mentions, emoticons by generic tags and to normalize the text to lower case.

In order to validate the results of the unsupervised labelling method and classification models, we ran three main experiment settings (see Section 7). Firstly, we validate the output of the unsupervised labelling method through the manual validation of a subset of the SS (ground truth set GT). In the second experiment setting, we evaluate the generated classification models taking as reference the ground truths generated for each domain. In addition, we run a test in the automotive domain using a gold standard (GS) dataset as reference during the validation phase.

Table 8 Statistics of the gold standard dataset and the datasets of each model in the literature related to the AP task based on business dimensions

Dataset	# Classes	# Users	# Tweets
GS	6	621	0
Humanizr	2	20,273	4,054,600
Demographer	2	208,000	208,000
M3	2	59,920	0
Full RepLab	10	7,000	4,200,000
RepLab mapped to SBI (cars)	3	1,620	0

The GS is a dataset built independently of the SS and manually labelled by human experts. We have built the GS in two steps. The first step is the automatic selection of a set of high-quality user descriptions. The second step is the manual selection and verification of the multi-level labels associated to them. In order to obtain high-quality profiles for authority classes, we organized a ranking based on the number of followers and verified user accounts. For the automotive domain, we also promote screen names associated to car brands or car components. All user accounts were checked to be active in Twitter. We also ensured that the dataset has a high heterogeneity degree by avoiding very similar descriptions to take part of the same class.

In a last experiment, we select the automotive domain to compare the resulting models against some state-of-the-art models related to some of the business dimensions we deal with. For this purpose, we again used the SS for training, but the evaluation was performed over the existing proposed datasets. Table 8 presents a summary of state-of-the-art and GS datasets.

Regarding the group perspective dimension (person-vs-organization), we evaluated our models over the curated dataset Humanizer (McCorriston et al., 2015), also used during models' evaluation of Demographer (Wood-Doughty et al., 2018) and M3 (Wang et al., 2019). We gathered the user profiles of Humanizr datasets, 20,273 accounts in total, of which 18,922 (93.3%) were still available in Humanizr shared repository as of May 2021.

We also evaluated our results against the RepLab dataset (Amigó et al., 2014) with the revision of tags presented in (Nebot et al., 2018). In this way, the profiles corresponding to RepLab classes "ngo", "celebrity","undecidable" and "sportsman" were removed from the evaluation because they do not map to our multi-dimensional profile model. We put together the RepLab classes "professional", "employee", "stockholder",

Table 9 Dimensions alignment with RepLab dataset

Business role dimension	RepLab Class	Influencers	All
Professional	company	69	203
	professional	238	855
	employee	6	10
	stockholder	0	0
	investor	0	0
Total professional		313	1068
Journalist	journalist	156	522
PublicService	public_institutions	1	30
Total		**470**	**1620**

Total users per dimension are shown in bold

"investor", and "company" under our "professional" class. We reported the evaluation scores for three class taxonomies, namely: "professional", "journalist" and "public_institutions", over the dataset corresponding to the automotive domain. Resulting alignments are shown in Table 9.

As an alternative, hypothesis verification can be achieved through visual data exploration (Keim & Ward, 2007), as well as, through automated techniques derived from statistics and machine learning. Figure 5 shows the results of applying T-SNE on two-dimensional transformed embeddings for the 6-class problem in the automotive domain. The embeddings were generated with fastText. The distribution of samples per class has been balanced (700 samples per class). From Fig. 5 we can see that most samples are clustered in their corresponding subgroups. If we analyse the figures from right to left, we first find the clusters of the three classes of category collective, while on the left the individual category profiles are grouped. The clusters of the classes 'professional.i. *' and 'journalist.i. *' are mixed due to the semantic similarity of the texts of their profiles. In general, the spatial representation of the embeddings is quite satisfactory.

6 Methods and Evaluation Metrics

In this section, we first describe all the methods and experimental settings used in our approach for the representation of texts and user profile classification. Then we describe the different classification tasks and finally the models evaluation measures are mentioned.

6.1 Methods

The method for the automatic construction of the SS makes use of fastText (Bojanowski et al., 2017) vectors for obtaining the semantic synonymy of words and bigrams (see Section 4.1). In this way, for each domain we create an unsupervised fastText model whose training corpus corresponds to the set of untagged user profiles

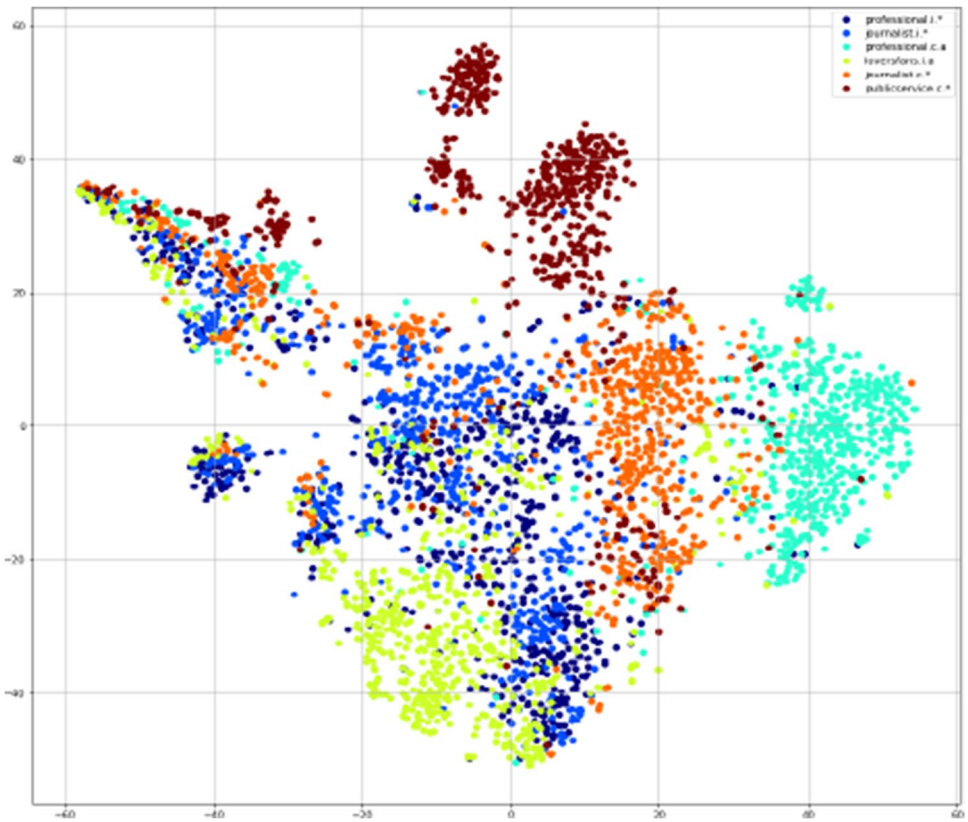


Fig. 5 2D T-SNE embeddings of user descriptions

(see Table 7). The trained word representations and training data set are then fed into the fastText classification model. To represent the embeddings of words and sentences we have selected fastText due to its simplicity, speed and efficiency. FastText is a library for efficient learning of word representations and sentence classification. One of its main contributions is that it uses the internal structure of the word to improve the representation of vectors. FastText enriches word vectors with sub-word (n-grams) information and word representation is learned by considering a window of left and right context words. The above allows the generation of embeddings for any word, even out-of-vocabulary words. This feature is very useful to represent embeddings of contents from social networks where text abbreviations are frequently used, new words arise, and spelling errors occur. Other methods like Word2Vec and Glove handle whole words but they can't easily handle words they haven't seen before. This fastText capability also allows capturing the underlying similarity between words, as for example can associate words that can be used in different contexts as "boxer" and "boxing". FastText offers similarity functions based on cosine distance between vectors. This similarity is computed for all words in the pre-trained vocabulary, allowing us to obtain the nearest neighbours of a given word and use functions to find analogies between words. For the classification task, multinomial logistic regression is used, where the sentence/document vector corresponds to the features. In (Joulin et al., 2016) refers that experiments show that fastText is often on par with deep learning classifiers in terms of accuracy, and many orders of magnitude faster for training and evaluation.

FastText Unsupervised Model for Word Representations After extensive testing of the model parameters, we set them as follows: CBoW with 100 dimensions, 10 epochs, minimal word occurrences of 5, negative sampling loss and number of negatives samples of 10, char n-gram between 5 and 8, and sampling threshold of 1e-05.

WordNet Super-Senses In order to find the most similar words to a given unigram, it is not enough to compute the nearest neighbours in the embeddings space. This function helps to find related but not synonymous words. To solve this problem, we propose the use of the semantic resource WordNet super-senses. WordNet lexical database (Miller, 1995) is the most commonly used resource for sense relation in English and many other languages. Content words (nouns, verbs, adjectives and adverbs) are grouped into sets of near-synonyms called synsets, each expressing a distinct concept. WordNet organize all word senses into lexicographic categories also called super-senses of which 26 are for nouns, 15 for verbs, 3 for adjectives and 1 for adverbs. In the context of our task, we define that two words are similar if their cosine similarity exceeds a given threshold but they must also meet the condition of belonging to the same semantic group (super-sense category in WordNet). To do this, we look up each word in the WordNet super-senses resource and check that they both represent the same semantic group. In this way we can validate a higher degree of similarity between words.

We have applied three kinds of classifiers for the experiments, namely fastText, RoBERTa (Liu et al., 2019) and a zero-shot classifier (Romera-Paredes & Torr, 2015; Zhang et al., 2019). We also used a majority class classifier as baseline model in our experiments.

One of the proposed classifiers is based on transfer learning. Transfer learning has attracted extensive attention in the scientific literature. It is a deep learning approach that uses existing relevant knowledge to solve new tasks in related fields (Peng et al., 2020). The resulting model is fine-tuned by a small dataset that is used to perform a specific task. For this study we have selected RoBERTa, a pre-trained model developed by Facebook AI based on Google's Bidirectional Encoder Representations from Transformers (BERT). RoBERTa has been shown better performance than BERT and LSTM models for Twitter text classification, even with small training datasets (Choudrie et al., 2021). RoBERTa is a robustly optimized BERT model (Devlin et al., 2019), whose pre-trained models have shown very effective for many downstream tasks like text classification. In the proposed experiments, we have just fine-tuned the pre-trained RoBERTa-base model with the SS dataset for author profile classification.

Zero-shot classification is aimed at categorizing data without using any training samples. In this paradigm, the user briefly describes each intended category, and the algorithm predicts for each instance its most likely category according to their semantics. This paradigm is well suited to our goal as we start with a brief description of each profile class, and then we aim to classify the incoming user descriptions into these categories.

The classification experiment settings are detailed as follows. For each domain, we trained a fastText classifier with the corresponding SS dataset, using the embeddings of the unsupervised fastText models. The empirical evaluation yielded the best results for the following parameters: minimal word occurrences of 2, char n-grams between 3 and 5, and 10 epochs. Then, we trained the following supervised models:

fastText 5-CV SS / RoBERTa 5-CV SS. A 5-fold stratified cross validation method was implemented for assessing a classifier trained and tested with the SS dataset.
fastText 5-CV GS / RoBERTa 5-CV GS. A 5-fold stratified cross validation strategy for assessing a classification model trained and tested with the GS dataset. It is a supervised classification model whose evaluation metrics serve as a reference to evaluate the models whose datasets were obtained in an unsupervised way.
fastText SS-GT / RoBERTa SS-GT. A classification model trained with the SS and tested with the GT.

fastText GT-GS / RoBERTa GT-GS. A classification model trained with the GT and tested with the GS.
fastText SS-GS / RoBERTa SS-GS. A classification model trained with the SS and tested with the GS.
Zero-Shot. We have used the new pipeline for zero-shot text classification included in the Hugging Face transformers package (Vaswani et al., 2017). Due to the computational cost of running this model on very large data sets, the evaluation was performed only over the gold standard dataset.

6.2 Classification Tasks

As the nature of the profiles is multi-level, we prepared four tasks and evaluated different predictive models for different class perspectives. The first model aims to classify profiles in the main perspectives shown in Table 2 (i.e., all levels expect the Domain role). This results into a 6-class problem.

The second model aims to classify only at the Business role level, becoming a 4-class classification problem (i.e., Journalist, Public Services, Professional and Lovers&Fans). In this way, the dataset labels are rewritten by keeping only the Business Role tag. For example, samples with tags 'journalist.c.*' and 'journalist.i. *' are rewritten to "journalist". This approach makes it possible to check whether there is a strong semantic relationship between the defined perspectives and to simplify the classification of new data sets.

The third and fourth models are binary classification problems. They are aimed to classify at the Group role level (Collective vs. Individual) and Domain Role level (On Domain vs. Off Domain) respectively.

In the case of Group role, it is worth mentioning that we checked whether the 6-class classification model improved its prediction, since the rest of tags could offer information on the group level. Although the results were good enough, they did not overcome those of the binary model.

6.3 Evaluation Measures

The measures selected to evaluate the results are micro and macro averaged F1-score, or the micro-F1 and macro-F1 scores respectively. Moreover, micro-F1 is also the classifier's overall accuracy, i.e., the proportion of correctly classified samples out of all the samples. The following always holds true for the micro-F1 case: micro-F1 = micro-precision = micro-recall = accuracy. The Area Under the Curve - Receiver Operating Characteristics (AUC-ROC) metric is also provided in a One-vs-Rest approach.

7 Evaluation Results

In this section, we test our different proposed models against current state of the art systems. There is no author classification system based on the proposed dimensions and levels for SBI; so, in a first tests stage we had used our own gold standard and ground truth datasets to evaluate our SS datasets and models for different classification tasks. In a second tests stage, we limited the evaluation to the classification perspectives of the only publicly available datasets of the state of the art.

7.1 Validation of the Unsupervised Labelling on the Three Domains

The problem of automatic tagging of Twitter profiles based on the detection of bigrams associated with different categories is a complex task. The interrelation between bigrams and unigrams associated with different classes makes it much more difficult to develop explicit rules to correctly identify the label of a profile. Therefore, human validation is required since it allows identifying the origin of potential errors and, if necessary, redefining class seed unigrams and conflict rules. Human validation is performed by a team of experts familiar with the domains and their business goals. Next, we describe the validation that have been assessed by the experts:

Cars Domain: We randomly selected 25 profiles per each of the 6 defined classes, resulting in 150 samples out of the 20,155 of the labelled profiles.
Tourism: We randomly selected 25 profiles per each of the 8 defined classes, resulting in 200 samples out of the 32,584 labelled profiles.
Medicine: We randomly selected 25 profiles per each of the 7 defined classes, resulting in 175 samples out of the 11,613 labelled profiles.

The three experts reviewed the same data individually and flagged the cases in which the assigned label was not correct. After checking the agreement in the categories assigned by the reviewers, the precision metric was used to evaluate the assessment.

The business dimensions proposed for the analysis are transversal to each domain. But within each domain the dimensions must be adjusted to their own categories or levels. For example, for the medicine domain (Aramburu et al., 2020), the category "professional.individual.medicine" represents all professionals in the health sector. On the other hand, the centers that offer health services are

Table 10 Statistics of the SS dataset construction for three different domains

Domain	Professional		Journalist		Services	Other relevant audience
	Individual	Collective	Individual	Collective		
Cars	U: 5 B:375 P:7963 C:1758	U: 8 B:89 P:4360 C:385	U: 4 B: 119 P: 2791 C:1485	U: 2 B: 78 P:1076 C:173	U: 5 B:46 P:1016 C:441	U: 3 B:47 P:2949 C:664
Tourism	U: 6 B: 16 P: 314 C: 116	U: 6 B: 20 P: 153 C:144	U: 1 B: 1275 P: 25,531 C: 1011	U: 2 B: 924 P: 2965 C:983	U: 6 B: 46 P: 3428 C:133	U: 2 B: 15 P: 193 C:126
Medicine	U: 6 B: 46 P: 1402 C: 109	-	U: 1 B: 48 P: 1487 C: 115	U: 2 B: 349 P: 4014 C: 175	U:2 B:142 P: 878 C:168	U:14 B:135 P:3832 C:194

Letter U represents the total of unigrams, the letter B the total of bigrams, the letter P the total of profiles and the letter C the total of conflicts

represented by a new higher-level category called "services.collective.medicine" that groups together both public and private services. In the case of the tourism domain, the category "professional.individual.tourism" represents all professionals in the tourism sector. While the category "professional.collective.tourism" represents companies within the sector. We have divided the "professional" dimension in the tourism sector into four subcategories in order to better identify the sets of associated profiles, namely: transportation, hotel, restaurant and entertainment. We have used these four new categories as part of the seed unigrams to identify the "tourism professional" dimension in the profiles. While to distinguish the individual and collective categories, the unigrams "professional" and "company" have been used, respectively.

Examples of bigrams of the class "professional.individual.tourism" are (tourism, professional) and (hotel, professional), whereas for "professional.collective.tourism" are (tourism, company) and (hotel, company).

Another interesting issue is that the initial unigrams represent categories at a very high level, like (tourism, professional), which allows carrying out a cold start of the models, without the need of having knowledge about the terminology used in the social network. Additionally, by adding seed unigrams specific to professional categories (e.g., waitress, hotelier, or stewardess) can improve the recall of the final model.

Table 10 shows the statistics of the obtained SS datasets for each domain. The table organizes the statistics of the results of different processes within the proposed method. The fields represent the dimensions of analysis shared by the proposed domains. In the table, the Professional and Journalist classes are represented up to the group level (collective and individual). On the other hand, the Services field represents the classes that are related to public and private services and will always be collective. Finally, the field "Other relevant audience" represents individuals that represent a type of audience interested in the domain. In the cases of the car and tourism domains, their audience is those categorized as Lovers&Fans of topics related to the domain. For the domain of medicine, we have grouped three categories of audience in this field: concerned (person interested in health issues, has survived an illness, etc.), student (a student of any school level) and religion (religious person). Table 11 shows the expert's revision results of the validation sets for each domain (i.e., precision metric). During the manual check, a few accounts were detected that are suspended on Twitter. In the statistics shown, the suspended accounts have been counted as labeling errors.

The statistics in Table 11 demonstrate the advantages of using the proposed semi-automatic method, since starting with a few seed unigrams we obtain a large number of labelled samples. The number of seed unigrams mainly depends on the heterogeneity of the profile class (e.g., concerned people in Medicine). The proportion of conflicts generated during tagging usually indicates the complexity of each class. Conflicts are also useful to redefine if necessary the seed unigrams as well as further conflict resolution rules. It is worth mentioning that the richer the conflict resolution rules the more variety is achieved in the final SS dataset.

7.2 Validation of the Models

In this section, we evaluate different classification models with the objective of validating the proposed method for AP. First, we prepare a model for the 6-class task based on the 3 proposed analysis perspectives (Business role, Group role and Domain role). To validate the proposed data processing steps for obtaining a valid SS, we evaluated a fastText model at 4 main phases

Table 11 Experts' assessment of the validation sets for each domain (precision)

Domain	Professional		Journalist		Services	Other relevant audience	Avg.
	Individual	Collective	Individual	Collective			
Cars	0.88	1	0.88	0.92	0.83	0.96	0.91
Tourism	0.84	1	0.96	0.96	0.84	0.84	0.91
Medicine	0.96	-	0.76	0.84	0.88	0.97	0.88
Avg.	0.89	1	0.87	0.91	0.85	0.92	

of the proposed method for different numbers of synonymy words (*K* parameter). The results showed that at each phase the accuracy of the resulting model was indeed increased.

For the cars domain experiment setting, we report the F1-micro and F1-macro scores for the different resulting models at each method phase for different *K* values, $K = \{1, 2, 3, 5, 10, 20, 30\}$. For each *K*, we test different models using fastText and RoBERTa, and the confidence intervals for 10 executions of each model were evaluated. We found no statistically difference between the different *K* values. To set a criterion for the model tests, we chose the value of $K = 5$, because its variance was slightly lower than the rest.

To validate the resulting classification models of the proposed method we run different experiment settings. Table 12 shows the evaluation, for each analysis domain, of fastText and RoBERTa models using a 5-fold stratified cross validation strategy trained with the corresponding SS, for the all tags classification task. The above evaluation was useful to check how close the prediction was to unsupervised labeling. These results are surely high because the train and test sets share textual patterns that are the bigrams and unigrams that generated them and this information is used by the classification models.

In order to check the predictive power of the model trained with the SS with respect to the GT, a model per domain was trained for predicting all the multidimensional tags. Table 13 shows the results for fastText and RoBERTa classifiers. Compared with the results of the manual validation of the GT subset (see Table 11), the trained models improve the scores of the labeling result, being very close to the decision of the human expert. The table also shows that RoBERTa outperforms fastText results only in the medicine domain. In the rest of the domains, both models have a similar performance.

The following experiments were performed only on the automotive domain. Tables 14 and 15 show the result scores, F1 and AUC-ROC scores respectively, for the models prepared for four different evaluation tasks, namely: all tags classification (6-class task), Business role (4-class task), Group role (2-class task) and Domain role (2-class task). To validate these models on an unseen set built independently of the SS, we ran a test using the GS dataset. In addition, any profile in the training set matching the GS was removed from the SS. The results are shown for the rows titled fastText SS-GS and RoBERTa SS-GS. To check if the results are good enough, we built different reference models trained with different ground-truth datasets (GT and GS) to be able to compare the prediction power of the trained models. If the results of SS dataset trained model are similar to the ground-truth dataset trained models, we can then claim the usefulness of the proposed unsupervised label method. Table 14 shows the results under the rows titled fastText 5-CV GS, RoBERTa 5-CV GS, fastText GT-GS, RoBERTa GT-GS, fastText GS-GT and RoBERTa GS-GT (refer to Section 6.1 to see the models details). Two baselines models are also presented in the table, the majority class and zero-shot classification. The best scores for the reference models have been marked in bold. On the other hand, the best results for the models trained with SS have been marked in bold and italics.

Notice that fastText and RoBERTa algorithms far exceed the baselines. FastText is an implementation that solves computing power problems, since it is a CPU efficient tool that allows models to be trained without the need for a GPU. On the other hand, BERT related models need the use of a GPU. In this sense, fastText models were trained in a server with 24 cores and 100Gb of RAM. The average time for training and predicting with fastText was 73 s. On the other

Table 12 Scores mean using a 5-fold cross validation stratified strategy over SS.

Model	Cars		Tourism		Medicine	
	Mean F1-μ	Mean F1-M	Mean F1-μ	Mean F1-M	Mean F1-μ	Mean F1-M
fastText 5-CV SS	0.985	0.983	0.995	0.983	0.977	0.970
RoBERTa 5-CV SS	**0.992**	**0.991**	**0.997**	**0.988**	**0.997**	**0.997**

The best results are in bold

Table 13 Models trained with full SS and validated over ground truth dataset

Model	Cars (6 classes)		Tourism (8 classes)		Medicine (7 classes)	
	Mean F1-μ	Mean F1-M	Mean F1-μ	Mean F1-M	Mean F1-μ	Mean F1-M
fastText SS-GT	**0.971**	**0.971**	**0.936**	**0.937**	0.968	0.968
RoBERTa SS-GT	0.963	0.962	0.930	0.933	**0.993**	**0.994**

The best results are in bold

Table 14 Evaluation of models using the GS and GT dataset. Results in italic correspond to best evaluations on the GS dataset

Model	ALL TAGS		Business role		Group role		Domain role	
	F1-μ	F1-M	F1-μ	F1-M	F1-μ	F1-M	F1-μ	F1-M
fastText 5-CV GS	0.888	0.886	0.893	0.890	0.965	0.965	0.915	0.902
RoBERTa 5-CV GS	**0.924**	**0.934**	**0.930**	**0.927**	0.977	0.977	**0.940**	**0.933**
fastText GT-GS	0.84	0.84	0.872	0.862	0.918	0.918	0.892	0.879
RoBERTa GT-GS	0.849	0.848	0.860	0.863	0.905	0.904	0.90	0.887
fastText GS-GT	0.869	0.868	0.883	0.863	**0.978**	**0.978**	0.920	0.908
RoBERTa GS-GT	0.883	0.880	0.869	0.860	0.971	0.971	0.912	0.90
fastText SS-GS	***0.881***	***0.882***	***0.877***	0.866	0.950	0.950	0.903	0.895
RoBERTa SS-GS	0.868	0.869	0.875	***0.871***	***0.959***	***0.958***	***0.919***	***0.912***
Majority class	0.181	0.051	0.346	0.128	0.501	0.338	0.659	0.397
Zeroshot GS	0.299	0.115	0.492	0.391	0.748	0.525	0.772	0.764

Bold refer to the best results along the column associated with models trained with datasets labeled by human experts (gold standard and ground truth datasets). Bold italics represent the best results achieved along the column for the models that were trained with the silver standard dataset

Table 15 AUC-ROC metric evaluation

Model/ AUC-ROC	ALL TAGS	Business role	Group role	Domain role
fastText 5-CV GS	0.935	0.927	0.966	0.910
RoBERTa 5-CV GS	**0.956**	**0.950**	0.977	**0.932**
fastText GT-GS	0.90	0.902	0.917	0.874
RoBERTa GT-GS	0.911	0.904	0.903	0.898
fastText GS-GT	0.917	0.895	**0.978**	0.890
RoBERTa GS-GT	0.926	0.894	0.970	0.88
fastText SS-GS	***0.933***	0.910	0.950	0.909
RoBERTa SS-GS	0.918	***0.919***	***0.958***	***0.924***

Bold refer to the best results along the column associated with models trained with datasets labeled by human experts (gold standard and ground truth datasets). Bold italics represent the best results achieved along the column for the models that were trained with the silver standard dataset

hand, the RoBERTa models were trained on a server with 8 Intel7 cores, 64Gb of RAM, and a 24Gb RTX3090 GPU. For the training of 20 epochs, it took 121 s on average to train and predict. Both models performed very similarly in terms of effectiveness, so we performed an unpaired t-test, which confirmed that the difference in the results is not statistically significant. However, fastText trains much faster and requires less computational resources than RoBERTa (e.g., use of GPU). Therefore, it is enough to use the fastText model for the AP task under the described conditions.

The validation results of the models under the 5-CV GS strategy offer information on the degree of difficulty of making predictions on the GS dataset. The results are high because the models perform the test on a small subset of the GS dataset in each iteration. Based on these results, we verify that the models trained with the SS are

Table 16 Results using the balanced SS dataset for training

Model	F1-μ	F1-M	Individual	Collective
Majority class	0.5	0.3333	1	0
fastText (6 classes)	**0.8691**	**0.7376**	**0.88**	0.75
fastText (2 classes)	0.8527	0.7234	0.86	0.79
RoBERTa (6 classes)	0.8057	0.6570	0.82	0.68
RoBERTa (2 classes)	0.6297	0.5481	0.59	**0.95**

The best results are in bold

Table 17 Results for the full SS dataset used during training

Model	F1-μ	F1-M	Individual	Collective
Majority class	0.8922	0.4715	1	0
fastText (6 classes)	**0.9033**	**0.7618**	**0.93**	0.63
fastText (2 classes)	0.7818	0.6547	0.78	0.82
RoBERTa (6 classes)	0.8712	0.7433	0.88	0.77
RoBERTa (2 classes)	0.8661	0.7498	0.87	**0.85**

The best results are in bold

not far from the best result of the reference models. On the other hand, models trained with SS outperform the models trained with the strategy GT-GS. The difference in the results may be due to the number of samples and words learned during the training. Also, comparing the strategies GS-GT with results on Table 13 for the automotive domain, the results of the models trained with the SS are better. The scores of the AUC-ROC metric (Table 15) confirm these results. As a conclusion, we can say that the proposed unsupervised labelling method is useful and effective.

7.3 Collective vs. Individual

We ran the following experiment to evaluate and compare our method for the task of distinguishing between individual and collective user profiles. For this purpose, we used the dataset given by (McCorriston et al., 2015) as gold standard. We evaluated the best models trained over the SS dataset for fastText and RoBERTa. In all experiments, the validation sets are the 100% of the data released by (McCorriston et al., 2015). The evaluation measures used were F1-micro, F1-macro and percentage of true positives per class. Table 16 shows the results of the evaluation of models trained with a balanced distribution of the classes of the silver standard dataset. Table 17 shows the results of the evaluation of trained models with the natural distribution of the SS classes. The exception is the Majority class method that works only with the class distribution.

According to the results of the previous tables, if we compare the percentage of hits per class, we can see that, for training with balanced classes, fastText performs much better than RoBERTa. However, when training with the full SS dataset, the best models were obtained for RoBERTa.

Table 18 shows the results reported by the state-of-the-art works for the prediction on the naturally distributed dataset. The M3 model presents the macro F1 value, while the other works report the accuracy value. The last two columns show the accuracy for each class.

While the state-of-the-art methods slightly outperforms our models in terms of global accuracy, all of them require significantly more data and features, as well as a manually curated dataset for training their models.

Comparing the results to (McCorriston et al., 2015), (Wood-Doughty et al., 2018) and (Wang et al., 2019), the best fastText model achieved 0.86 of accuracy when trained with a balanced SS subset. For the naturally distributed SS, fastText achieved the best score with 0.90 of accuracy, but predictions are very similar to the majority class baseline.

Finally, the models that best tradeoff offer for the two classes are those of RoBERTa. RoBERTa model achieved 0.85 of accuracy for the collective class compared to the best model of M3 with 0.807. Thus, our approach achieved the best score for the collective class, which is a major source of error in the other approaches in the literature (Wang et al., 2019).

7.4 RepLab Evaluation

In this section, we discuss the evaluation results on the RepLab 2014 for author profiling dataset. It is important to emphasize that RepLab's author profiling task and ours are different. Firstly, some of the dimensions of RepLab are transversal to the categories of our SS dataset. For example, in RepLab, the classes journalist and professional contains both collective and individuals. Besides, RepLab consider the class company separately from professional. For the sake of simplification, we placed under our professional user profile the following RepLab classes: professional, stockholder, investor, employee and company. Besides, we discarded the Undecidable and NGO classes, as they do not fit well into our multidimensional model. We also removed the "sportsmen" class because it involves several classes of our profile model: athletes (professional.i.*), sport business (professional.c.*) and sports newspapers (journalist.c.*).

It must be noticed that the evaluation results are not comparable to those published in (Amigó et al., 2014) and (Nebot et al., 2018), since the models were trained on different types of data and features, with different class configurations. In our case, we trained the models with the user descriptions of the silver standard, and then evaluated with the RepLab user accounts mapped to our multi-dimensional model.

Table 18 Results of the best models in the literature compared to our method

Model	Collective	Individual
Humanizr McCorriston et al. (2015)	0.586	0.982
Demographer Wood-Doughty et al. (2018)	0.644	0.973
M3 Full, Multilingual Wang et al. (2019)	0.807	**0.986**
Our method (RoBERTa)	**0.850**	0.870

The best results are in bold

Table 19 Results over the *influencer's* subset of the RepLab 2014 automotive domain

Model	F1-μ	F1-M	P-μ	P-M
fastText $K=5$	0.628	0.415	0.627	0.428
fastText $K=20$	**0.679**	**0.447**	**0.679**	0.440
fastText $K=30$	**0.679**	0.438	**0.679**	0.448
RoBERTa $K=5$	0.589	0.421	0.589	**0.468**
Majority class baseline	0.530	0.231	0.530	0.177

The highest result along each column is highlighted in bold

Table 20 Results over the full dataset of the RepLab 2014 automotive domain

Model	F1-μ	F1-M	P-μ	P-M
fastText $K=5$	0.596	0.364	0.596	0.336
fastText $K=20$	**0.6617**	**0.5026**	**0.6617**	0.4859
fastText $K=30$	0.6531	0.4930	0.6531	0.4771
RoBERTa $K=5$	0.5494	0.4366	0.5494	**0.4929**
Majority class baseline	0.4302	0.2005	0.4302	0.1434

The highest result along each column is highlighted in bold

Tables 19 and 20 show the results obtained for the influencer subset and the overall RepLab dataset. Notice that the overall performance of our method is poorer than in the other datasets. After inspecting the RepLab dataset in depth, we realized that the categorization criteria of users in RepLab differed notably from ours. Experts in RepLab were more focused on the users' "intentions" rather than their reported description. For example, a well-known TV media was tagged as NGO because its posts were usually related to regular campaigns related to social activities. As a result, author profiling in RepLab is a very hard task, and no approach in the literature achieved good results for this dataset (Nebot et al., 2018).

7.5 Discussion

The experiments allowed validating the starting hypothesis, demonstrating that the labeling of the most frequent bi-grams related to the predefined topics by class captures the semantics of the profiles associated with that class. From the manual labeling of a small seed of unigrams and bi-grams within a data stream, it was possible to build a large and representative dataset of the intended classes. According to the results, we can conclude that the vocabulary in the profiles is sufficient to learn signs of the different proposed business dimensions. An interesting finding was that our method significantly outperforms the Zero-shot text classification method used as baseline.

Compared to other approaches in the literature, our method performs well for the Group role perspective. Unfortunately, we only found one dataset for the Business role perspective: RepLab 2014. Our method performs poorly in this dataset similarly to other approaches in the literature. Our main conclusion is that this dataset was designed with different criteria, which are not correlated to user descriptions.

After performing an error analysis on the GS results, we found that most errors come from Lovers&Fans individuals misclassified as Professionals. This is because many professionals also express their interests and hobbies in their descriptions along with their professional occupation. Similar errors occur between the Journalist and Professional collective classes, where some specialized magazines can be identified as professional media. Indeed, as their goals and descriptions are very similar, the decision to classify a specialized magazine as Journalist or Professional is quite subjective. Therefore, analysts should decide which class best fits these conflicting samples accordingly to their tasks.

7.6 Limitations

Our approach has the following limitations. First, the work focuses on the multiclass classification problem, therefore only a single label can be assigned to each user profile. However, many Twitter users often list in their description all the roles they take on (e.g. different occupations, hobbies) and have taken on (e.g. previous occupations) in their life. When different roles are listed in the same profile, it becomes very difficult to identify the primary role. In (Sloan et al., 2015), when multiple occupations appear, the first mentioned one is selected. In RepLab, experts were mainly focused on classifying the user's intention (through

the analysis of their posts) rather than on the own profile description (see Section 7.4 and 7.5). Consequently, discrepancies and difficulties arise when only a single label (per dimension) is assigned to complex profiles. In our method, we make use of the conflict resolution rules to make decision on the final single label per profile dimension. A second limitation is that the proposed method can only be applied to well-formalized business domains, where prior knowledge is available regarding their organization, hierarchies, and relationships. On the other hand, there are other more abstract domains whose classes are more complex to model only with bigrams, such as the identification of influencers or the classification of hackers, bots, spammers and haters in social media. These types of tasks usually require to analyze more information like the interactions between users, posts text, network structure, etc. Finally, biases have not been treated in depth in this work. The biases can be conditioned by many factors, including the training datasets, the pre-trained word embeddings or the algorithm itself. In this work, we alleviate the problem of bias in the Dataset Debugger phase by applying conflict resolution rules. However, more research needs to be done for detecting possible bias in the labelled datasets as well as reducing its impact in the trained models.

8 Conclusions

This paper proposes a methodology for author profiling (AP) in Twitter based on social business intelligence roles. The method allows the unsupervised construction of a labelled dataset that serves as input to different text classification tasks. Unlike most approaches on AP for social networks, we do not analyze nor mine the tweets contents for inferring the author's roles. Instead, we automatically build a training dataset from unlabelled user descriptions by making use of the multidimensional user profile knowledge model provided by the analysts.

Existing AP approaches are time-consuming and resource demanding. These solutions use ad-hoc machine learning methods based on large manually labelled training datasets, which are hard to build and difficult to be extended to other analysis tasks or contexts.

From the theoretical point of view, this study makes important contributions to the way AP is performed within social business intelligence (SBI) projects. Firstly, we propose a novel method for automatically building training datasets aimed at classifying user profiles in dynamic scenarios. Our method relies on semantic knowledge represented by ontologies provided by the analysts when starting a SBI project, from which basic linguistic information is extracted to identify candidate unlabelled user profiles. In this way, the generated training data are directly linked to the concepts represented in the knowledge multidimensional model (i.e., users' roles). Thus, we can check consistency and conflicts in the training data. As a result, the proposed method contributes to the current state of the art with an integrated view of ontologies and predictive models for AP in social networks. Moreover, the method is able to deal in dynamic scenarios where the semantic knowledge needs to be updated to consider new roles or rejecting others. In this case, the method just builds a new training dataset and new predictive models for the updated knowledge. Another implication of this paper is that the use of the user profiles, instead of their posts and/or metrics, is enough for characterizing their business roles. Previous methods based on posts and metrics obtained very poor results due to the nature of social network contents, which is very redundant, usually shared, and heterogeneous. Empirical results on different domains demonstrated the usefulness and effectiveness of the proposed method.

One of the greatest difficulties that analysts face when addressing machine learning tasks is the difficulty to obtain a good enough training dataset. In a SBI project, user roles and categories are defined on demand, so we cannot assume the existence of labeled datasets for them. Thus, **the main practical implication** of this work is that our approach can help to identify relevant user groups with minimal human intervention. This approach helps to save time, to reduce human effort and economic resources to obtain the predictive models for user profiling. This results in a dynamic framework where target profiles can be easily adapted to new on-demand needs. Furthermore, our proposed method achieved competitive results in all tested datasets, including those of the state-of-the-art that treated some SBI perspectives like Group and Business roles.

The multidimensional AP approach responds to the needs for analysis based on dynamic dimensions that arise in social media. Multidimensional AP can help informational systems to characterize the audience of published popular topics and news. As an example, we used AP in the domain of intelligent health surveillance (Aramburu et al., 2020) for detecting fake topics related to health. In this scenario, fake topics were dominated by non-expert audiences, whereas true topics involved relevant users with professional roles. Data quality can be also greatly benefit from AP as it is a direct quality indicator of the data sources. Preliminary results presented in (Aramburu et al., 2021) showed that user roles are relevant for quantifying the quality of Twitter data streams. In both studies, we manually built the user groups, which was a difficult and time-consuming task. With the proposed method, the user groups are defined much more quickly and at a much larger extent.

Future work aims at improving the classification of certain user categories that were more difficult to predict due to mixed contents. On the other hand, the study of the social metrics that best model each of the defined classes would allow us to build tighter predictive models enabling

us to classify a larger proportion of users, especially when the user has no description field. Finally, we plan to perform some contrasting study in order to evaluate how well the user descriptions fit with the contents they generate, as well as to find correlations between descriptions and other aspects like psychological traits and emotions of users.

Regarding the limitations section, we plan to evaluate different methods for bias correction like AFLite (Le Bras et al., 2020) to improve the variability of the resulting training datasets. We also plan to study multi-output classifiers for detecting secondary roles for complex profiles as well as assigning to them a probability distribution instead of a unique category.

Acknowledgements This research has been funded by the Spanish Ministry of Industry and Commerce grant number PDC2021-121097-I00, and by the pre-doctoral grant of the Universitat Jaume I with reference PREDOC/2017/28.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Declarations

Conflict of Interest Authors have no conflict of interests to declare.

References

Amigó, E., Carrillo-de-Albornoz, E., Chugur, I., Corujo, A., Gonzalo, J., Meij, E., de Rijke, M., & Spina, D. (2014). Overview of RepLab 2014: author profiling and reputation dimensions for online reputation management. In E. Kanoulas, M. Lupu, P. Clough, M. Sanderson, M. Hall, A. Hanbury, & E. Toms (Eds.), *Information Access evaluation. Multilinguality, Multimodality, and Interaction* (8685 vol.). Springer. Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-319-11382-1_24

Aramburu, M. J., Berlanga, R., & Lanza-Cruz, I. (2021). Quality management in social business intelligence projects. In *Proceedings of the 23rd International Conference on Enterprise Information Systems - Volume 1: ICEIS* (pp. 320–327). https://doi.org/10.5220/0010495703200327

Aramburu, M. J., Berlanga, R., & Lanza-Cruz, I. (2020). Social media multidimensional analysis for intelligent health surveillance. *International Journal of Environmental Research and Public Health, 17*, 2289. https://doi.org/10.3390/ijerph17072289

Aswani, R., Kar, A. K., & Vigneswara Ilavarasan, P. (2018). Detection of spammers in twitter marketing: a hybrid approach using social media analytics and bio inspired computing. *Information Systems Frontiers, 20*, 515–530. https://doi.org/10.1007/s10796-017-9805-8

Berlanga, R., García-Moya, L., Nebot, V., Aramburu, M. J., Sanz, I., & Llidó, D. M. (2015). SLOD-BI: an open data infrastructure for enabling social business intelligence. *International Journal of Data Warehousing and Mining (IJDWM), 11*(4), 1–28. https://doi.org/10.4018/ijdwm.2015100101

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146. arXiv:1607.04606v2.

Carrillo-de-Albornoz, J., Gonzalo, J., & Amigó, E. (2019). RepLab: an evaluation campaign for online monitoring systems. In N. Ferro & C. Peters (Eds.), *Information Retrieval Evaluation in a Changing World*. The Information Retrieval Series, vol 41. Springer. https://doi.org/10.1007/978-3-030-22948-1_20

Cervero, R., Rosso, P., & Pasi, G. (2021). Profiling fake news spreaders: personality and visual information Matter. In E. Métais, F. Meziane, H. Horacek, & E. Kapetanios (Eds.), *Lecture notes in Computer Science* (p. 12801). Springer. Natural Language Processing and Information Systems. https://doi.org/10.1007/978-3-030-80599-9_31

Choudrie, J., Patil, S., Kotecha, K., et al. (2021). Applying and understanding an advanced, novel deep learning approach: a covid 19, text based, emotions analysis study. *Information Systems Frontiers, 23*, 1431–1465. https://doi.org/10.1007/s10796-021-10152-6

Ciaramita, M., & Johnson, M. (2003). Supersense tagging of unknown nouns in WordNet. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (pp 168-175). EMNLP 2003. https://aclanthology.org/W03-1022

Daelemans, W., et al. (2019). Overview of PAN 2019: bots and gender profiling, celebrity profiling, cross-domain authorship attribution and style change detection. In F. Crestani, et al. (Eds.), *Lecture notes in Computer Science, vol11696, experimental IR meets multilinguality, multimodality, and Interaction*. Springer. https://doi.org/10.1007/978-3-030-28577-7_30

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics, NAACL*.

Gallinucci, E., Golfarelli, M., & Rizzi, S. (2015). Advanced topic modeling for social business intelligence. *Information Systems, 53*, 87–106. https://doi.org/10.1016/j.is.2015.04.005

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv:1607.01759v3. https://doi.org/10.48550/arXiv.1607.01759

Keim, D., & Ward, M. (2007). Visualization. In M. Berthold & D. J. Hand (Eds) *Intelligent data analysis*. Springer.

Kim, A., Miano, T., Chew, R., Eggers, M., & Nonnemaker, J. (2017). Classification of twitter users who tweet about E-cigarettes. *JMIR Public Health and Surveillance*, *3*. https://doi.org/10.2196/publichealth.8060

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: the definitive guide to dimensional modeling*. Wiley.

Kumar, U., Reganti, A. N., Maheshwari, T., et al. (2018). Inducing personalities and values from language use in social network communities. *Information Systems Frontiers, 20*, 1219–1240. https://doi.org/10.1007/s10796-017-9793-8

Lanza-Cruz, I., Berlanga, R., & Aramburu, M. J. (2018). Modeling analytical streams for social business intelligence. *Informatics, 5*, MDPI.

Le Bras, R., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M., Sabharwal, A., & Choi, Y. (2020, November). Adversarial filters of dataset biases. *In International Conference on Machine Learning* (pp. 1078–1088). PMLR.

Li, H., Mukherjee, A., Liu, B., Kornfield, R., & Emery, S. L. (2014). Detecting campaign promoters on twitter using markov random fields. *2014 IEEE International Conference on Data Mining*, 290–299.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: a robustly optimized BERT pretraining approach. *ArXiv, abs/1907.11692*.

López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H. J., Pineda, L. V., & Stamatatos, E. (2015). Discriminative subprofile-specific representations for author profiling in social media. *Knowledge Based Systems, 89*, 134–147.

López-Santillán, R., Montes-y-Gómez, M., González-Gurrola, L. C., Alonso, G. R., & Prieto-Ordaz, O. (2020). Richer document embeddings for author profiling tasks based on a heuristic search. *Information Processing & Management, 57*, 102227.

Markov, I., Gómez-Adorno, H., Posadas-Durán, J. P., Sidorov, G., & Gelbukh, A. (2017). Author profiling with Doc2vec neural network-based document embeddings. In O. Pichardo-Lagunas, & S. Miranda-Jiménez (Eds.), *Advances in Soft Computing*. *Lecture notes in Computer Science* (10062 vol.). Springer. https://doi.org/10.1007/978-3-319-62428-0_9

McCorriston, J., Jurgens, D., & Ruths, D. (2015). Organizations are users too: characterizing and detecting the presence of organizations on Twitter. *International AAAI Conference on Web and Social Media*, *ICWSM*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Proceedings of a meeting held December 5-8, 2013. *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013* (pp. 3111–3119). Lake Tahoe, Nevada, United States. https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications Of The Acm, 38*, 39–41.

Mishra, P., Tredici, M. D., Yannakoudakis, H., & Shutova, E. (2018). Author profiling for abuse detection. *International Conference on Computational Linguistics*, *COLING*.

Nebot, V., Pardo, F. M., Berlanga, R., & Rosso, P. (2018). Identifying and classifying influencers in Twitter only with textual information. In M. Silberztein, F. Atigui, E. Kornyshova, E. Métais & F. Meziane. (Eds.), *Lecture Notes in Computer Science*, vol 10859, *Natural Language Processing and Information Systems*. NLDB 2018. Springer. https://doi.org/10.1007/978-3-319-91947-8_3

Ouni, S., Fkih, F., & Omri, M. (2021). Toward a new approach to author profiling based on the extraction of statistical features. *Social Network Analysis and Mining, 11*, 1–16. https://doi.org/10.1007/s13278-021-00768-6

Pardo, F.M., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., & Stein, B. (2016). Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations. In Series CEUR Workshop Proceedings vol.1609. *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation Forum* (pp. 750-784). Evora, Portugal. CEURWS.org. http://ceur-ws.org/Vol-1609/16090750.pdf

Peng, D., Wang, Y., Liu, C., et al. (2020). TL-NER: a transfer learning model for chinese named entity recognition. *Information Systems Frontiers, 22*, 1291–1304. https://doi.org/10.1007/s10796-019-09932-y

Pennacchiotti, M., & Popescu, A. (2011). A machine learning approach to Twitter user classification. *International AAAI Conference on Weblogs and Social Media*, *ICWSM*.

Potthast, M., Rosso, P., Stamatatos, E., & Stein, B. (2019). A decade of shared tasks in digital text forensics at PAN. In L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, & D. Hiemstra, (Eds.), *Lecture Notes in Computer Science*, vol 11438. Springer. https://doi.org/10.1007/978-3-030-15719-7_39

Rangel, F., Sarracén, G. L., Chulvi, B., Fersini, E., & Rosso, P. (2021). *Profiling hate speech spreaders on Twitter Task at PAN 2021*. CLEF, CEUR-WS.org.

Rodríguez-Vidal, J., Carrillo-de-Albornoz, J., Gonzalo, J., & Plaza, L. (2021). Authority and priority signals in automatic summary generation for online reputation management. *Journal of the Association for Information Science and Technology, 72*, 583–594. https://doi.org/10.1002/asi.24425

Romera-Paredes, B., & Torr, P. H. (2015). An embarrassingly simple approach to zero-shot learning. *International Conference on Machine Learning, ICML.*

Schler, J., Koppel, M., Argamon, S. E., & Pennebaker, J. W. (2006). Effects of age and gender on blogging. *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*.

Schlicht, I. B., & Magnossão de Paula, A., F. (2021). Unified and multilingual author profiling for detecting haters. *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, 2936*, 1837–1845. https://dblp.org/rec/conf/clef/SchlichtP21.bib

Sloan, L., Morgan, J., Burnap, P., & Williams, M. (2015). Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PLoS One, 10*(3), e0115545.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems,* (pp. 5998–6008). ArXiv, abs/1706.03762.

W3C Semantic Web (2012). *SKOS. Simple knowledge organization System*. https://www.w3.org/2004/02/skos/. Accessed 25 May 2021

W3C Recommendation OWL (2004). Web ontology language guide. https://www.w3.org/TR/owl-guide/. Accessed 14 Sept 2021

Wang, Z., Hale, S. A., Adelani, D., Grabowicz, P. A., Hartmann, T., Flöck, F., & Jurgens, D. (2019). Demographic inference and representative population estimates from multilingual social media data. *The World Wide Web Conference*.

Weren, E. R., Kauer, A. U., Mizusaki, L., Moreira, V. P., Oliveira, J. P., & Wives, L. (2014). Examining multiple features for author profiling. *J Inf Data Manag, 5*, 266–279.

Wood-Doughty, Z., Mahajan, P., & Dredze, M. (2018). Johns Hopkins or johnny-hopkins: classifying individuals versus organizations on Twitter. *Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, PEOPLES@NAACL-HTL.*

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing [Review Article]. *IEEE Computational Intelligence Magazine, 13*, 55–75. https://doi.org/10.1109/MCI.2018.2840738

Zanakis, S. H., & Evans, J. R. (1981). Heuristic "Optimization": why, when, and how to use it. *Interfaces, 11*, 84–91.

Zhang, J., Lertvittayakumjorn, P., & Guo, Y. (2019). Integrating semantic knowledge to tackle zero-shot text classification. *NAACL-HLT ArXiv, abs/1903.12626*.

Zhang, X., Tsang, A., Yue, W. T., et al. (2015). The classification of hackers by knowledge exchange behaviors. *Information Systems Frontiers, 17*, 1239–1251. https://doi.org/10.1007/s10796-015-9567-0

Indira Lanza-Cruz is a researcher and lecturer in the Department of Computer Languages and Systems at the Universitat Jaume I, Spain, and is a member of the Temporal Knowledge Bases Group. She researches in the fields of Data Science and Intelligent Systems for Knowledge Management applied to Social Business Intelligence. She is a Ph.D. student in Computer Science at the same university. She holds a Master's Degree in Intelligent Systems from the Universitat Jaume I in 2016 and a Master's Degree in Computer Engineering and Networks from the University of Seville in 2015. She graduated as a Computer Engineer and Professor of Computer Science at the "Instituto Superior Politécnico José Antonio Echeverría" (CUJAE) in Cuba in 2006. Since then, she has had the opportunity to combine academic research with the development of solutions and services for the company. Her current lines of research are aimed at developing predictive models for the creation of strategic corporate indicators from social and organizational data.

Rafael Berlanga Llavori is a full-time Professor at the Universitat Jaume I, where he has been teaching and researching for more than 20 years. He has been leading the Temporal Knowledge Bases research group since then. His main research lines are mainly focused on the area of intelligent information processing systems, especially in the acquisition and management of knowledge from textual sources and its application to business intelligence scenarios. He formerly contributed to the area of temporal reasoning, and afterwards to the development of novel knowledgebased techniques within the semantic web field. More recently, he also achieved relevant results in applying statistical language models to sentiment analysis and social network analysis. He also designed novel text mining techniques for biomedical applications. Most of this research was transferred to the spin-off SemanticBots, from which he is a founding partner. These research results were achieved in the context of eleven doctoral theses, and were continuously funded through competitive projects of the national and European R&D programmes and several R&D contracts with TI companies.

María José Aramburu is full professor of Computer Science at Universitat Jaume I, where she is full time member of TKBG (Temporal Knowledge Bases Group) since 1997. She obtained a B.S. degree in Computer Science from Universidad Politécnica de Valencia in 1991, and a Ph.D. from the School of Computer Science of the University of Birmingham (UK) in 1998. Her main research interests include integration and exploitation of semistructured data, advanced business intelligence applications, Semantic Web, and quality management of social media data.

Chapter 5

Quality Management in Social Business Intelligence Projects

Publications

Berlanga, Rafael ; Lanza Cruz, Indira Lázara and Aramburu, María José. "Quality Indicators for Social Business Intelligence," 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, 2019, pp. 229-236, doi: 10.1109/SNAMS.2019.8931862.

Aramburu, María José; Berlanga, Rafael and Lanza Cruz, Indira. (2023). A Data Quality Multidimensional Model for Social Media Analysis. Business & Information Systems Engineering, DOI: 10.1007/s12599-023-00840-9 (In Press). (Q1)

Quality Indicators for Social Business Intelligence

Rafael Berlanga
Department de Llenguatges i Sistemes Informàtics
Universitat Jaume I
Castelló de la Plana, Spain
berlanga@uji.es

Indira Lanza-Cruz
Department de Llenguatges i Sistemes Informàtics
Universitat Jaume I
Castelló de la Plana, Spain
lanza@uji.es

María José Aramburu
Department d'Enginyeria i Ciència dels Computadors
Universitat Jaume I
Castelló de la Plana, Spain
aramburu@uji.es

***Abstract*—The main purpose of Social Business Intelligence is to help companies in making decisions by performing multidimensional analysis of the relevant information disseminated on social networks. Although data quality is a general issue in SBI, few approaches have aimed at assessing it for any data collection, being this a context dependent task. In this paper, we define a quality indicator as a metric that serves to assess the overall quality of a collection, and that integrates the measures obtained by several quality criteria applied to filter the posts relevant for a SBI project. The selection of the best quality criteria to include in each quality indicator is a complex task that requires a deep understanding of both the context and objectives of analysis. In this paper, we propose a new methodology to design quality indicators for SBI projects whose quality criteria consider contents coherence and data provenance. Thus, for the context defined by an objective of analysis, this methodology helps users to find the quality criteria that best suit both the users and the available data, and then integrate them into a valid quality indicator.**

Keywords—data quality, business intelligence, social media, business indicators

I. INTRODUCTION

Social networks have become a new source of useful information for companies, helping them, among others, to know the opinions of their customers, to analyse the trends of their market, and to discover new business opportunities [1] [2]. Social media constitutes a fundamental part of the information ecosystem, and there has been a growing interest in the development of solutions for social data analysis from the commercial and scientific perspectives.

The main purpose of Social Business Intelligence (SBI) is to help companies in making decisions by performing multidimensional analysis of the relevant information disseminated on social networks. However, today businesses mainly use social networks to produce a group of social metrics [3] [4] that are analysed in an isolated way. Rarely, companies integrate social metrics with other business measures to calculate and analyse key performance indicators. The fundamental cause of this underutilization is the lack of trust that companies have in this kind of data, since, coming from social networks, they do not have control over their origin, contents and quality.

In an environment as open and out of control as Twitter, or any other social network, it is difficult to find the right data for a SBI project. As in any other Business Intelligence application, the subject of analysis is described, as exactly as possible, by using natural language expressions. Then, a proper group of keywords are chosen to program the retrieval mechanisms. However, as experience shows, the resulting collection use to be noisy, incomplete or biased. Most times, it will include posts generated with very different purposes, as well as posts without a true relation with the subject of analysis. Their specific contents may even be counterproductive and produce false conclusions, due to the misinformation and the noise they generate. This is a data quality problem that requires executing data cleansing operations on posts retrieved from social networks before extracting any social metrics from them. Thus, the execution of a SBI project must start by assessing the quality of the available data and, when needed, improving it until the established standard in reached [2].

Although data quality is a general issue in SBI [2] [5], few approaches have aimed at assessing this data aspect for large collections of posts. Most approaches just apply a series of ad-hoc quality rules over posts (e.g., tweets with more than three retweets, users with more than 100 followers, and so on) in order to filter those that will be analysed. In addition, many works aim at analysing concrete events such as a catastrophe or a terrorist attack where the main issue is to score tweets credibility [6] [7].

Evaluating the quality of social data is a context dependent task [8], and each SBI project will require the definition of the best quality indicators for the source data. We define a quality indicator as a metric that serves to assess the overall quality of a collection and that integrates the measures obtained by the various quality criteria. With this approach, as in [8], several quality metrics will serve to filter the posts that are relevant for a SBI project. However, the selection of the best quality criteria and metrics, and how to combine them into a quality indicator, is a complex task that requires a deep understanding of both the context and objectives of analysis.

In this paper, we propose a new methodology to design quality indicators for SBI projects. After analysing the quality attributes of tweets, we have identified that the main quality criteria for SBI analysis are contents coherence and data provenance, two aspects not treated in the literature. Thus, our methodology relies on two foundations. Firstly, considering the collection of tweets of an analysis context, it is possible to model the language of the context and measure the contents coherence of each tweet, as well as each user profile description. These measures, plus other useful metrics from users and tweets, such as the number of favourites, replies, repeats, and followers, are the main quality criteria to consider by our methodology. Secondly, a classification of the different user profile categories that participate in the context of analysis helps to identify the best quality criteria for each kind of user. In this way, given an objective of analysis in a specific context, with this methodology it is possible to find the quality criteria that best suit both its users and the available data, and then integrate them into a quality indicator valid for that SBI project.

The main idea behind this methodology is that the users with a Twitter profile description in high correlation to the subject of analysis and showing a coherent activity in their

posts will have higher quality than the rest of users. These users will constitute the group of relevant users of a SBI project and will serve as reference to build quality indicators. By finding the quality criteria that produce good measures for the relevant users of an analysis context it is possible to identify the best quality metrics for that SBI project. Our method also helps to integrate these metrics in order to build effective quality indicators.

The rest of the paper is organised as follows. Section II reviews SBI approaches and methodologies. Section III specifies the analysis dimensions that are involved in the assessment of social data quality. A methodology for measuring quality in SBI projects is described in Section IV, and some results and conclusions are presented in sections V and VI respectively.

II. Social Business Intelligence Approaches

SBI is an emerging discipline that combines corporate data with content generated by users on social media with the aim of improving decision making in the company [2] [4]. Social networks can be analysed from different user perspectives such as contents, relationships and behaviour, becoming an abundant source of information on opinions, interests, needs and attitudes of users. The challenge lies in the efficient management of information from social networks considered as Big Data, characterized by an immense amount and variety of unstructured and noisy data, which change at a high speed.

This section provides a review of the current state of art of Social Business Intelligence that distinguishes between methodologies for the development of new applications and the quality assessment of social media data. It also depicts the main issues and novelties of our approach.

A. SBI Methodologies

Despite the growing interest in the development of SBI applications in the industry, in the scientific literature there have been very few approaches that establish a methodology for their design and implementation. In general, SBI requires highly integrated multidisciplinary research [9] and here we highlight the main approaches of methodologies and architectures.

The work in [10] proposes an iterative methodology for the design and maintenance of SBI applications, establishing an ordering for the common tasks executed during social media analysis. It emphasizes the need for agile and effective support during the maintenance of the infrastructure, due to the dynamism of the user generated contents and the changes in the environment. The main tasks proposed are iterative and can be optimized, they are organized as follows: macro-analysis (definition of the scope of the project and the questions it will solve), ontology design, source selection, semantic enrichment, crawling design, ETL and OLAP design, and finally, the execution and testing. In turn, the authors propose the development of an architecture where the information resulting from the analysis is stored within a data mart in the form of multidimensional cubes that can be exploited by OLAP techniques. The basic problem is that all the information is stored in a historical repository that requires large storage resources. The task of analysing the relevance of the contents generated during the crawling process relies on a manual labelling that distinguishes between in-topic and off-topic clips (textual data). In a later stage, the recovered documents are filtered based on search criteria that do not guarantee the credibility and quality of the source. On the other hand, when analysing large volumes of data this process can be unsustainable.

More in accordance with the current needs for SBI and Big Data processing, a new approach focuses on the speed and immediacy of information, processing data in streaming and incorporating batch analysis processes to obtain knowledge models. With the learned models it is possible to apply inductive processing algorithms on the data stream and to favour the semantic enrichment of the data [11] [12]. Only the information elements that are needed for the knowledge models are stored, optimizing memory usage. However, so far, there are few solutions for SBI, and analysis tasks are mainly oriented towards event detection [13] and recommendation systems [12].

The SLOD-BI project [14] is an infrastructure for linked open data that lays the foundation for the implementation of SBI. It offers mechanisms for the extraction, linking and publication of social data in the form of RDF triplets modelled as multidimensional stars. Providing access to large open knowledge resources and multidimensional analytical models to define efficient methods of data extraction and analysis, the project proposes the combination of cognitive models with statistical language models to infer useful information from the texts generated by the users. A semantic meta-structure of multidimensional analytical patterns (user facts, social facts and dimensions) is proposed to model social data in a way that facilitates the integration with corporate data stored in traditional repositories and data warehouses. In [15] SLOD-BI is extended with both a methodology and a framework oriented towards the formalization of social indicators and performance measures to support decision-making. This proposal allows social measures exploration and aggregation over dynamic multidimensional contexts for on-demand objectives.

Currently, the main challenges that SBI projects have to face are the high dynamicity of both the elements implied in the analytics and the analytical requests, as well as, the high percentage of noisy data. In this sense, [16] proposes an architecture and a methodology to model analytical streams for SBI that relies on both linked data and multidimensional modelling. The architecture eases the cleaning and semantic enrichment of data, whereas the methodology serves to shape the data for analysis purposes. The adoption of semantics facilitates the development, validation and follow-up of workflows. Thus, instead of storing the semantically enriched facts, they can be generated and processed on the fly. The solution corresponds to a Kappa streaming architecture that consists of two stages: a long-term stage for keeping recent historical information as a data stream of long duration (to collect data for inductive processing tasks) and a short-term stage with some workflows for generating the required analysis data in real-time.

All these approaches (i.e., [10] [11] [12] [14]) have in common the semantic enrichment of the social data in order to increase the processing capabilities of the collection. Entity resolution and semantic enrichment do not only provide a good context for understanding post contents, but also transform them into meaningful data with a representation easier to process by analysis tools. In this sense, semantic enrichment processes serve to enhance the quality of social data collections. However, they cannot measure the quality and filter the posts that a SBI project require.

B. Social Data Quality Assessment

When implementing a SBI project, it is critical to assess the quantity and quality of the available data, since without an amount large enough of valuable data, any implementation of a BI-oriented application will fail to [17] [18]. However, as shown by a recent review paper [5], few approaches to business social media analytics depict pre-analytics processing activities, and the task of assessing data quality is left out of consideration in all revised works.

In [8], the profiling and supervision of the quality of the data are considered as primary concerns of the Big Data processing cycle. With this methodology, the user is in charge of analysing the quality metrics provided by the system. In their solution, a quality management module includes an interface that enables the end-user to configure the quality policies of a company. The module also includes a quality evaluator for visualizing the values for the metrics of the quality attributes of each police. Similarly, to the quality indicators of our approach, quality policies combine several metrics to measure the quality of a collection. However, the work in [8] does not address the problem of finding and integrating the best quality metrics to assess data quality in a specific SBI project.

In spite of the great interest that the analysis of the quality of social data arouses in the scientific community, there are very few related works within the scope of SBI. The evaluation of the quality of contents published on microblogging platforms has focused mainly on post retrieval operations. Searching for posts related to a topic [19] [20]; filtering posts based on their credibility and quality [21] [22]; detection of events and disasters [13] [23] [24] [25]; analysis of feelings, political and consumer opinions [14] [26] [27]; and detection of influencers [28] [29], are some example applications. Other applications aimed at the detection of spammers, bots and advertising campaigns, have proposed intelligent analysis techniques for social metrics [30] [31] [32].

In the literature, quality measures are defined at post and user level. At post level, there are a large number of metrics covering the characteristics of the text (e.g., grammar, contents and semantics), together with the metrics specific to microblogs that reflect their social impact (e.g., number of comments and retweets). On the other hand, at user level, there are activity metrics to assess the relevance (e.g., account age and number of posts) and popularity (e.g., number of followers, likes and mentions) of issuers.

Most of the quality attributes applicable for social media data classified in [8] can easily be measured by means of the contents and metadata of Twitter posts. Among them, relevance (measured as post content keywords), believability (measured with author attributes like followers count or registration age), popularity (measured as number of readers or re-tweets) and timeliness (i.e., tweet date). Corroboration (i.e., the number of data sets where the issue has been recognized) and validity (i.e., likelihood of data validity for its purpose) are other quality attributes included in this classification. Although, during data extraction, these two attributes cannot be measured for a single data element, they are considered important when evaluating several data sets in the data analysis phase. The paper does not provide any methods to measure them.

C. Our Approach

In contrast to previous approaches, in our approach, we consider contents coherence together with the origin of the social data (i.e., data provenance) as fundamental quality attributes. So far, the relevance of a post was ensured when it contained one or more retrieval keywords. As it is recognised, keywords are imprecise, and using them to build a complete collection produces the retrieval of many useless posts. With our approach, by measuring the coherence of the posts contents with respect to the language model of the analysis context, it is possible to filter noisy posts and to improve the quality of the collection.

On the other hand, Twitter user profiles can constitute a source of valuable information about posts and issuers that previous approaches have not taken into consideration. From our point of view, measuring the coherence of user profile descriptions with respect to the language model of the analysis context also helps to identify the relevant posts of an SBI project. The challenge lies in developing efficient natural language processing mechanisms to identify semantic and syntactic patterns within the texts of a fully open environment such as the Twitter social network. Here, tweets are composed of very short texts, written with a style that is informal and full of hashtags and abbreviations, and produced by a large range of different kinds of users, including the noise produced by fake users and bots.

These two new quality attributes can be applied together with other commonly used metrics to assess the relevance, believability and popularity of each post and its issuer. However, the idea of integrating several quality measures into a single quality indicator is completely new. Up to our knowledge, there are not previous works that have proposed a method to define quality indicators for social data [33]. The quality policies proposed by [8] allow users to describe the relevant quality attributes and their importance for a SBI project. However, this work does not propose a method to build indicators from a set of metrics with the purpose of assessing and analysing the overall quality of a collection of posts from different perspectives (e.g. types of social network users, temporal evolution or geographical distribution). In the next sections, we present a semi-automatic method that, using a ranking of relevant users as a reference, helps to create quality indicators and social media data collections.

III. Analysis Dimensions for Quality Assessment

Some of the most powerful mechanisms for modelling analysis dimensions are the categories of customer segmentation, which vary depending on the business context. For example, demographic data and behavioural styles are usually applied to model the categories of the customer dimension. In the case of social media data, the user that writes a post can represent an individual person or a company, and it can play a role that depends on the context of analysis. Thus, our approach models the different business roles that the social media users of an SBI project can take as a way of providing a full range of perspectives of analysis.

More specifically, we propose four disjoint categories for the "user business role" dimension, named as follows: Professional, Journalist, Public Service and Lovers & Fans. The first three categories fall within the services sector, whereas the category Lovers & Fans comprises general users who use social networks by their own interest and could generate posts relevant to the context of analysis. Below, we

describe these four categories and give examples of Twitter user profile descriptions associated with each category in the context of an automotive-related SBI project.

A. *Professional*

The "Professional" category is related to a subset of activities in the services sector. In turn, we divide this class into two subcategories.

"Professional on domain", which covers official companies, small dealers or professional groups within the analysis domain that use Twitter to promote their services. The following is an example of a professional in the automotive sector:

```
since 1921, pep boys has been the top
automotive aftermarket chain w/ quality
auto repair & car parts in 800 locations
across 35 states & puerto rico
```

"Professional others" represents professionals who do not belong to the domain of analysis. The following example shows an instance of this category:

```
sales director @adeogroup providing
ecommerse and web solutions including
responsive web design to a diverse range
of clients visit {lnk}
```

B. *Journalist*

This category includes users dedicated to the publication of news, such us journalists, newspapers and magazines.

```
wews newschannel5 is on your side with
breaking news & weather updates from
northeast ohio
```

C. *Public service*

This category represents organizations dedicated to public services such as government agencies, emergency and security services.

```
york regional police official twitter . in
case of emergency dial 911/non-emergency
call 1-866-876-5423 . account is not
monitored 24/7
```

D. *Lovers & Fans*

This category represents people or groups that do not use social networks with a professional interest, but as a means of learning, fun and entertainment. They are usually consumers of information instead of emitters. This category represents the highest percentage of individual users in social networks and constitutes the main source of noise for many analysis contexts.

```
book lover , gamer , cat wrangler ,
autoimmune fighter . tweets : {lnk}"
```

IV. A Methodology for Measuring Quality in SBI Projects

As in our previous work, we rely on a linked open data infrastructure where social network data is continuously stored as social BI facts, called SLOD-BI [14]. In this paper, we aim at directly deriving the quality indicators from the metrics associated to users and posts within the infrastructure. These quality indicators will be stored and used by the own infrastructure to processes and filter social data before their analysis. Figure 1 shows an example of a quality indicator derived from the post and user facts, and relying on the business role dimension.

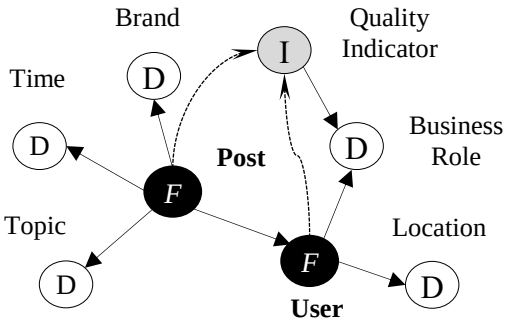


Fig. 1. Quality indicator (I) derived from Social BI facts. As in traditional BI, social facts (F) consist of metrics and dimensions (D). Metrics have been omitted in the figure.

Quality indicators must be adapted to the needs of the analysis at hand. As we adopt multi-dimensional models for social BI analysis, quality indicators can be also associated to some specific analysis dimensions. For example, in Fig. 1, the quality indicator is associated to the business role dimension, so that we can get different quality criteria depending on this perspective.

A quality indicator is formally expressed as a formula over fact metrics, which calculates the quality score of each fact according to the selected perspective. This quality score can be then applied to filter the data to be analysed.

The proposed method can be summarised in the following steps:

1) First, we establish a reference dataset of relevant users associated to the dimension of the quality indicator we aim at (e.g., business role in Figure 1).

2) Then, we select a series of metrics associated to user and post facts that are deemed as relevant indicators of quality. Some of these metrics are directly taken from the infrastructure (e.g., followers, retweets, etc.) whereas others are derived by processing both the post contents (e.g., length of messages) and the user descriptions (e.g., size of the description).

3) A quality score is calculated for each selected fact metric and each dimension perspective.

4) The formula of the quality indicator is derived by a simple linear combination of the fact metrics. The weights of the linear combination are directly obtained from the quality scores of Step 3.

5) Finally, the obtained quality indicator is applied over each incoming post/user fact to decide whether it is included in the analysis or not.

We describe in detail these steps in turn.

Step 1: Reference dataset of relevant users

In this work, we consider relevant users those that publish reliable messages according to the analysis task at hand. Useful analyses should rely on high quality data, which mainly comes from reliable users. In this way, we assume that the provenance of posts is a primary quality criterion. Another relevant criterion for quality is the coherence of the published contents and their sources. Since we cannot know all relevant users publishing in a data stream, we should rely on a small reference set of relevant users to assess the impact of existing metrics on their quality level. The main idea is then to predict

a quality formula that can rank incoming user/post facts similarly to the reference set of relevant users.

For selecting relevant users we could make use of existing resources such as analytical platforms, like Socialbakers®, or existing datasets, like RepLab 2014 [34]. However, these resources are limited to a few vertical domains and mainly aim at identifying influencers from different perspectives. Indeed, users included in these resources usually have a very large number of followers. For this reason, in this paper we propose a different way to obtain reference relevant users according to their descriptions.

The proposed method consists in manually labelling the top most frequent word bigrams of the user profile descriptions. In Twitter these descriptions are included along with the tweet metadata, so this information can be easily obtained from the incoming data stream. Bigrams are labelled with the dimension values associated to the quality indicator. For example, following the schema of Figure 1, bigrams are labelled with the values of the business roles (e.g., journalist, professional and so on). Finally, user descriptions containing labelled bigrams with the same label (i.e. they have associated a unique dimension value) are considered relevant users for that label. Table I shows an example of labelled bigrams and the number of relevant users selected for each label.

Step 2: Selection of fact metrics

After analysing the different methods proposed in the literature for social data quality, we selected those metrics widely adopted for filtering posts. Then, we defined some new metrics to measure the coherence of texts in the posts and user descriptions with respect to the overall vocabulary of the stream. These metrics are organized into two levels: users and posts. For example, user quality is usually associated to metrics like number of followers, account age, and interaction metrics with other users. Posts metrics are related to metrics associated to the users' interactions with the post (e.g., likes, retweets, etc.) as well as metrics derived from the post text like sentiment indicators, semantics, and different lexical features.

Tables II and III show the selected metrics for users and posts respectively in our preliminary experiments. Concerning the user-level metrics (Table II), "interactions to post" is the number of interactions over posts performed by the user (e.g., retweets, likes, etc.), whereas "interactions from users" is the number of interactions received from other users. Description coherence is the entropy of the language model of the user description with respect to the overall language model of the stream. The lower the entropy the better correlated is the description with the domain. The account age is directly associated to the user identifier number (i.e., the lower the older). "Posts on domain" is the number of tweets published by the user in the analytical data stream. The rest of user-level metrics are provided as metadata by the incoming data stream.

With respect to the post-level metrics (Table III), "repetitions" is the number of times the same message has been published in the stream without being retweets. "Text coherence" is the entropy of the language model derived from the post text with respect to the overall language model of the domain. The next metrics are directly taken from the incoming data stream. Finally, the last four metrics are facts and sentiment indicators extracted from the post text [14].

Step 3: Impact of metrics

To measure the impact of a metric in the data quality, we propose a novel method based on Information Retrieval evaluation. Our hypothesis is that a metric will have a high impact in data quality if the ranking of items produced by that metric promotes the relevant users associated to a given dimension.

A direct way to measure this impact is to use de Mean Average Precision (MAP) metric [35]. A high MAP value indicates that relevant items are mostly placed at the top positions of the ranking. However, it heavily depends on the number of relevant items (the larger this number the higher the scores). As we use different datasets of relevant users for each dimension, we need to normalize MAP scores. In this work we propose to use the relative change with respect to a uniform distribution of relevant items.

$$MAP = \frac{\sum_{k=0}^{N} pre(k) \cdot rel(k)}{R}$$

$$MAP_{relative} = \frac{(MAP - N/R)}{MAP}$$

Where N is the total number of items in the ranking and R is the number of relevant items. Function *pre* returns the precision at position k in the ranking. Function *rel* returns 1 if the item at position k is relevant and 0 otherwise.

Notice that a value of $MAP_{relative}$ near to 0 indicates a low impact of the metric to rank relevant items. Notice also that large negative values indicate that the ranking should be reversed to have a positive impact in the promotion of relevant items.

Since metrics can be associated to either users or posts, depending on the target metric we will rank either users or posts to calculate the corresponding MAP score. Additionally, a MAP score can be calculated for each dimension value (labels) by considering only the corresponding subset of relevant items associated to it.

Step 4: Quality Indicators

The last step consists in applying the quality scores of each metric to obtain the corresponding quality indicators. In this paper we propose a simple linear combination of the metrics where impact scores act as weights. More specifically, given a fact F with metrics (m_1, ... , m_M), the quality indicator for dimension value Di is the following formula:

$$I_{Di}(F) = \sum_{k=0}^{M} \alpha_k^{Di} \cdot norm(F[m_k])$$

Here, the quality scores of each metric and dimension value are represented by the coefficients α_k^{Di}. To properly combine the metrics, we normalize and scale them with the function *norm* as described in turn.

As previously mentioned, when the quality factor of a metric is negative, the ranking must be inverted. For this purpose, we apply the complement with respect to the maximum value of the metric (i.e., max(m_k) - F[m_k]). Additionally, we must multiply by -1 to make the corresponding factor positive.

The function *norm* also applies a logarithm transformation to all the metrics that counts things (e.g., retweets, statuses,

etc.), since all these metrics follow a power law distribution. Finally, all metrics are normalized in the range [0, 100].

Step 5: Quality Data Filtering

Quality indicators can be applied in different ways to filter the incoming posts before analysing them. The most straightforward way is to firstly apply a threshold over the indicators at user-level, and then a second threshold over the indicators at post-level. Additionally, we can select one of the perspectives of the dimension associated to the quality indicator (e.g., the business role to be focused on).

V. RESULTS

For demonstrating the usefulness of the previous metrics, we have chosen a long-term stream of tweets related to the automotive domain. This stream has been active from 2015 until now and has served as basis of several studies about Social BI [14] [16]. The stream is generated with a series of keywords representing different car models and brands. It currently contains 1.930.617 tweets, written in both Spanish (456.059) and English (1.474.558).

Table I shows the different categories contained in the user dimension for performing different analysis tasks. The second row in this table indicates the number of relevant users for each class. An external indicator of the relevance of these users is the ratio of verified accounts, which is of 0.9% for the whole stream and 3.3% for the selected relevant users.

TABLE I. EXAMPLES OF TOP FREQUENT BIGRAMS ASSIGNED TO THE DIFFERENT USER CATEGORIES

Journalists (J)	·news information ·motoring news ·auto news
5173	
Lovers & Fans (L&F)	·sports fanatic ·love cars ·auto enthusiast
4572	
Professional on domain (P.D)	·used cars ·cars service ·car parts
4441	
Professional others (P.O)	·community manager ·project manager ·writer photographer
3286	
Public Service (PS)	·call emergency ·crime call ·report call
193	

Table II shows the impact weights obtained for the different user categories by applying the MAP scores and the relevant users associated to them. In this table, MAP scores have been transformed into normalized weights. Cells marked with asterisks correspond to negative MAP scores (see Step 4), and therefore their metrics should be inverted when contributing to the indicator. As can be noticed, different categories show quite different weights which leads to different quality indicators. Notice also that most categories promote the coherence of the user description with respect to the domain.

TABLE II. USER-LEVEL QUALITY SCORES

METRICS	**User business roles**				
	J	***L & F***	***P.D***	***P.O***	***PS***
Interactions to posts	0.01	0.09	0.09	0.05	0.04
Interactions from users	0.09	0.07*	0.12	0.04	**0.26**
Description coherence	0.16	**0.23**	**0.24**	**0.29**	0.11
Account age	0.05	0.13*	0.18	0.02	0.04
Posts on domain	0.19	0.08	0.07*	0.10	0.22
Followers	0.12	0.18	0.04*	0.17	0.04*
Friends	**0.24**	0.17	0.07*	0.24	0.20
Listed users	0.13	0.06	0.18*	0.10*	0.08*
Published posts	0.01	0.09	0.09	0.05	0.04

Table III shows the impact weights for the post-level metrics. We use the same conventions than in Table II. As for the user-level metrics, the different categories show different weight configurations. It can be noticed that the most relevant metrics correspond to relevant users of each perspective. Thus, journalist and public services promote tweets with a high number of interactions, whereas professional categories promote tweets with sentiment data. Notice also that automotive-related professionals (P.D) have a greater weight in the text coherence metric than other professionals (P.O).

TABLE III. POST-LEVEL QUALITY SCORES

METRICS	**User business roles**				
	J	***L & F***	***P.D***	***P.O***	***PS***
Repetitions	0.06*	0.02*	0.15*	0.03*	0.12*
Text Coherence	0.06*	0.09*	**0.21**	0.03*	**0.16***
Replies	0.02	0.10	0.01*	0.09	0.01*
Retweets	**0.21**	0.08	0.02*	0.09	**0.18**
Favourites	**0.22**	0.11	0.05*	0.09	0.14
Sentiment Score	0.04	0.12	**0.21**	**0.24**	0.14*
Sentiment facts	0.13	0.15	0.05	0.15	0.10
Opinion expressions	0.14	**0.17**	0.16	**0.16**	0.10
Feature expressions	0.12	**0.16**	0.13	0.12	0.06

Finally, once quality indicators are built, we need to set the thresholds that will be applied for data quality filtering. If no perspective is selected, we assume that the thresholds are applied to the best-scored perspective at user level. As

previously mentioned, metrics are normalized in the range [0, 100] and therefore final indicators will be also in that range.

When applying the best-scored perspective, the very dominant role is "Professional on domain", which mainly covers all the advertisements and promotions generated by agents of the automotive business. Threshold setting is performed by exhaustive exploration, looking for a trade-off between relevant user coverage and the size of the filtered dataset.

Alternatively, we can select the perspective that best fits our analysis. For example, we can select the journalist perspective if we want to analyse the events associated to the different brands. Threshold setting for each perspective can be different, as shown in Table IV. In this table, we fix the filter ratio to 20% to find out their corresponding thresholds. Journalist role gets the lowest thresholds, which indicates that quality scores generated for this perspective are usually lower than in the other ones because its profile is much more difficult to cope with.

TABLE IV. THRESOLD SETTINGS PER PERSPECTIVE

	J	**L&F**	**P.D**	**P.O**	**PS**
User-level threshold	>25	>25	>40	>25	>20
Post-level threshold	>5	>20	>20	>20	>20

[a.] filtering ratio fixed to 20%

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a novel methodology to semi-automatically define quality indicators for social BI. It relies on the selection of a ranking of relevant users associated to specific analysis dimensions. Taking this ranking as reference, the method automatically calculates the impact of each metric in the quality indicators applied to filter social data before analysis.

Preliminary results show the viability of the approach for a large stream of tweets in the automotive domain. The resulting quality indicators show that there is not a unique formula for all the analysis tasks but different configurations depending on the kind of contents we want to analyse.

Future work will be mainly focused on a comprehensive evaluation of existing social data metrics to make the obtained quality indicators much more effective. We will also analyse semantic-based methods to automate as much as possible the selection of relevant users for specific dimensions. Finally, further experiments must be done to measure the noise reduction produced by the proposed quality indicators.

ACKNOWLEDGMENT

This research has been funded by the Spanish Ministry of Industry and Commerce grant number TIN2017-88805-R and by the pre-doctoral grant of the Universitat Jaume I with reference PREDOC/2017/28.

REFERENCES

[1] Akter, S., Bhattacharya, M., Fosso Wamba, S. and Aditya, S. "How does Social Media Analytics Create Value?" in Journal of Organizational and End User Computing 28, 1-9, 2016.

[2] Ruhi, U. "Social Media Analytics as a BI Practice: Current Landscape and Future Prospects" in Journal of Internet Social Networking and Virtual Communities. 1-12, 2014.

[3] Keegan, B. and Rowley, J. "Evaluation and decision-making in social media marketing" in Management Decision. 55, pp. 15-31, 2017.

[4] Lee, I. "Social media analytics for enterprises: Typology, methods, and processes" in Business Horizons. 61.2. 199-210, , 2018.

[5] Holsapple, C., Hsiao, S. and Pakath, R. "Business social media analytics: Characterization and conceptual framework" in Decision Support Systems. 110, 2018.

[6] Gupta, A., Kumaraguru, P., Castillo, C. and Meier, P. "TweetCred: Real-Time Credibility Assessment of Content on Twitter" in Proceedings of the 6th International Conference on Social Informatics. 228-243, 2014.

[7] O'Donovan, J., Kang, B., Meyer, G., Höllerer, T. and Adalii, S. "Credibility in Context: An Analysis of Feature Distributions in Twitter" in International Conference on Privacy, Security, Risk and Trust, pp. 293-301, 2012.

[8] Immonen, A., Pääkkönen, P. and Ovaska, E. "Evaluating the Quality of Social Media Data in Big Data Architecture" in IEEE Access. 3, 2015.

[9] Zeng, D., Chen, H. and Lusch, R. "Social Media Analytics and Intelligence" in IEEE Intelligent Systems, Vol. 25, n. 6, pp. 13-16, 2010.

[10] Francia, M. Golfarelli, M.and Rizzi. S. "A methodology for social BI" in Proceedings of the 18th International Database Engineering & Applications Symposium, ACM, 2014.

[11] Barbieri D. et. Al., "Deductive and Inductive Stream Reasoning for Semantic Social Media Analytics" in IEEE Intelligent Systems, pp. 32-41, 2010.

[12] Nadal, S. et al. "A software reference architecture for semantic-aware Big Data systems" in Information & Software Technology 90, pp. 75-92, 2017.

[13] Zhang, C. et al. "GeoBurst: Real-Time Local Event Detection in Geo-Tagged Tweet Streams", In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, Pisa, 2016.

[14] Berlanga, R. et al. "SLOD-BI: An Open Data Infrastructure for Enabling Social Business Intelligence" in International Journal on Data Warehousing and Data Mining, vol. 11, 4, pp. 1-28, 2015.

[15] Lanza-Cruz, I. and Berlanga, R. "Defining Dynamic Indicators for Social Network Analysis: A Case Study in the Automotive Domain using Twitter" in the 10th International Joint Conference on Knowledge Discovery and Information Retrieval, pp. 219-226, 2018.

[16] Lanza-Cruz, I., Berlanga, R. and Aramburu, MJ. "Modeling Analytical Streams for Social Business Intelligence" in Informatics 5(3), 2018.

[17] Czernek, A. "Social Measurement Depends on Data Quantity and Quality" in Millward Brown Dynamic Logic, 2018. http://www.millwardbrown.com/docs/default-source/insight-documents/points-of-view/Millward_Brown_POV_Social_Measurement_Depends_on_Data_Quantity_and_Quality.pdf (accessed 5 September 2019).

[18] Inmon, B. "Data Lake Architecture: Designing the Data Lake and Avoiding the Garbage Dump". Technics Publications, 2016.

[19] Massoudi, K., Tsagkias, M., de Rijke, M. and Weerkamp, W. "Incorporating Query Expansion and Quality Indicators in Searching Microblog Posts" in Advances in Information Retrieval - 33rd European Conference on IR Research, pp. 18-21, 2011.

[20] Xie, W., Zhu, F., Jiang, . Lim, P. and Wang, K. "TopicSketch: Real-Time Bursty Topic Detection from Twitter" in IEEE Transactions on Knowledge and Data Engineering , pp. 2216 - 2229, 2016.

[21] Momeni , E., Tao K. and Haslhofer, B. "Identification of Useful User Comments in Social Media: A Case Study on Flickr Commons" in Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 1-10, 2013.

[22] Chen, W. et al. "A study on real-time low-quality content detection on Twitter from the users'perspective" in PLoS ONE, vol. 12, 8, https://doi.org/10.1371/journal.pone.0182487, 2017.

[23] Feng, W. et al. "STREAMCUBE: Hierarchical spatio-temporal hashtag clustering for event exploration over the Twitter stream" in Proceedings of the 2015 IEEE 31st International Conference on Data Engineering, pp. 13–17, 2015.

[24] Zhou, X. and Chen, L. "Event detection over twitter social media streams" in The VLDB Journal, pp. 381–400, 2014.

[25] Zubiaga, A., Spina, D., Martínez R. and Fresno, V. "Real - time classification of Twitter trends" in Journal of the Association for Information Science and Technology, pp. 462‑473, 2015.

[26] Liu, X. et al. "A Text Cube Approach to Human, Social and Cultural Behavior in the Twitter Stream," in International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, 2013.

[27] Rosenthal, S. Farra, N. and Nakov, P. "Sentiment Analysis in Twitter" in Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), Vancouver, Canada, 2017.

[28] Rodríguez - Vidal, J., Gonzalo, J., Plaza, L. and Anaya, H. " Automatic detection of influencers in social networks: Authority versus domain signals" in Journal of the Association for Information Science and Technology, vol. 70,7, pp. 675-684, 2019.

[29] Mahalakshmi, G. S., Koquilamballe, K. and Sendhilkumar, S. "Influential Detection in Twitter Using Tweet Quality Analysis" in Second International Conference on Recent Trends and Challenges in Computational Models, pp. 315-319, 2017.

[30] Miller, Z. et al. "Twitter spammer detection using data stream clustering" in Information Sciences, pp. 64-73, 2014.

[31] Varol, O. et al. "Online Human-Bot Interactions: Detection, Estimation, and Characterization" in Social and Information Networks, 2017.

[32] Li, H. et al. "Detecting Campaign Promoters on Twitter using Markov Random Fields" in theIEEE International Conference on Data Mining, Shenzhen, 2014.

[33] Taleb, I., Serhani, M. and Dssouli, R. "Big data quality: A survey" in Proc. IEEE Int. Congr. Big Data, pp. 166–173, 2018.

[34] Amigó, E. et al. "Overview of RepLab 2014: Author Profiling and Reputation Dimensions for Online Reputation Management" in Kanoulas E. et al. (eds) Information Access Evaluation. Multilinguality, Multimodality, and Interaction, 2014.

[35] Manning, C., Raghavan, P. and Schütze, H. "Introduction to Information Retrieval". Cambridge University Press, 2008.

Bus Inf Syst Eng
https://doi.org/10.1007/s12599-023-00840-9

RESEARCH PAPER

A Data Quality Multidimensional Model for Social Media Analysis

María José Aramburu · Rafael Berlanga · Indira Lanza-Cruz

Received: 28 November 2022 / Accepted: 12 September 2023

Abstract Social media platforms have become a new source of useful information for companies. Ensuring the business value of social media first requires an analysis of the quality of the relevant data and then the development of practical business intelligence solutions. This paper aims at building high-quality datasets for social business intelligence (SoBI). The proposed method offers an integrated and dynamic approach to identify the relevant quality metrics for each analysis domain. This method employs a novel multidimensional data model for the construction of cubes with impact measures for various quality metrics. In this model, quality metrics and indicators are organized in two main axes. The first one concerns the kind of facts to be extracted, namely: posts, users, and topics. The second axis refers to the quality perspectives to be assessed, namely: credibility, reputation, usefulness, and completeness. Additionally, quality cubes include a user-role dimension so that quality metrics can be evaluated in terms of the user business roles. To demonstrate the usefulness of this approach, the authors have applied their method to two separate domains: automotive business and natural disasters management. Results show that the trade-off between quantity and quality for social media data is focused on a small percentage of relevant users. Thus, data filtering can be easily performed by simply ranking the posts according to the quality metrics identified with the proposed method. As far as the authors know, this is the first approach that integrates both the extraction of analytical facts and the assessment of social media data quality in the same framework.

Keywords Data quality · Social media data · Business intelligence · Text analytics

1 Introduction

Social media has emerged as a valuable source of information for companies, enabling them to understand customer opinions, analyze market trends, and uncover new business opportunities, among other benefits (Ruhi 2014). Although social media contents are highly heterogeneous and difficult to manage, they can produce meaningful business information for decision-makers. The research presented here focuses on data quality management of social media data collections for Business Intelligence applications.

Business intelligence (BI) is the process of collecting, storing and analyzing data from business operations to assist organizations in becoming data-driven (Sabherwal and Becerra-Fernandez 2013). Although BI tools are primarily powered by operational data sources (i.e., OLTP data), they also allow business users to access heterogeneous types of data from historical/current, structured/unstructured, internal/external sources. BI user practices range from analytics and reporting to data mining and predictive analytics. BI platforms rely on data warehouses

Accepted after one revision by Óscar Pastor.

M. J. Aramburu (✉)
Department de'Enginyeria i Ciència dels Computadors, Universitat Jaume I, 12071 Castelló de la Plana, Spain
e-mail: aramburu@uji.es

R. Berlanga · I. Lanza-Cruz
Department de Llenguatges i Sistemes Informàtics, Universitat Jaume I, 12071 Castelló de la Plana, Spain
e-mail: berlanga@uji.es

I. Lanza-Cruz
e-mail: lanza@uji.es

Published online: 10 November 2023

for storing their reference information. More specifically, a traditional data warehouse aggregates operational data into multidimensional data structures applied by online analytical processing (OLAP) engines to execute data-intensive queries. Although BI tools are primarily descriptive in summarizing historical and present data evolution, data analytics modules can also become part of modern BI systems, enhancing them with statistical tools as well as artificial intelligence and machine learning capabilities. These tools enable deeper business insights producing business information ranging from descriptive to predictive, prescriptive, and self-explanatory (Gröger 2021).

Social Business Intelligence (SoBI) is defined in Gallinucci et al. (2015) as the discipline that aims at combining business with social media data to form a corporate data warehouse which lets decision-makers enhance their business needs based on the trends and moods perceived from the environment. Until now, companies have used social networks mainly for marketing purposes (Păvăloaia et al. 2020). In fact, SoBI tools are often applied by marketing departments to monitor the performance of their social media activities with metrics like number of likes, followers, or replies (Keegan and Rowley 2017; Lee 2018) as well as the feelings and concerns of their customers (Choi et al. 2020). However, social media data have many more uses and applications for businesses and industries (Gioti 2018), and integrating social media metrics with corporate data can help to produce better strategic indicators to drive companies forward (García-Moya et al. 2013; Ruhi 2014; Stieglitz et al. 2014).

Social media analytics (SMA) is another large family of related applications that can be defined as the ability to gather and find some meaning in social media data to aid business decisions and measure the performance of social media actions based on those decisions (Ruhi 2014). Dealing with social media data, these data analytics tools also form part of modern SoBI platforms as they can be applied to help business decisions (Holsapple et al. 2018). In a general business setting, SMA is focused on statistical and machine learning tools that apply correlation, regression, and classification, together with sentiment extraction, to transform social media data into meaningful information for business purposes (Stieglitz et al. 2014). SMA has many real-world applications and has been widely applied by the research community to solve different types of problems related to business management (Stieglitz et al. 2014; Zachlod et al. 2022).

In this paper, we use SoBI and SMA as a unified term and treat social media big data analytics as a related field. However, to emphasize our perspective of integrating social media data into a BI environment, we mainly use the term SoBI for the remainder of the paper. All these systems have in common that, to produce valuable insights, they require feeding with collections of social media data of good quality with respect to their analysis objectives. However, building good quality collections is difficult because social media posts consist of unstructured texts with a high level of semantic heterogeneity. Fake posts, jokes, bots, and misinformation are often mixed with serious user-generated contents. In addition, the range of users participating in social networks is also diverse and their posts serve very different purposes. In a business environment, it is possible to find anonymous customers who publish opinions about a brand (e.g., offers, products or services), employees of the company who generate ads for marketing purposes, and other professionals posting contents somehow related to the brand (e.g., journalists, professional or customer associations, influencers, etc.).

Current approaches build collections of social media data by translating a subject of analysis into a set of retrieval keywords (i.e., topics, usernames, and hashtags). These keywords are then applied to filter one or several social networks (e.g., Twitter, Facebook, etc.), generating in this way a stream of potentially relevant posts with different degrees of quality for the analysis objectives (Holsapple et al. 2018; Arolfo et al. 2022). Most times, in the large volume of data retrieved, there are many posts apparently related to the subject of analysis but turn out because of their origin, intention or specific contents, to be useless and to produce noise or misunderstandings (Aramburu et al. 2021). These posts do not add any value to the analysis tasks and may even be counterproductive, due to the misinformation and the noise they produce. For example, in our experiments, when attempting to gather customer opinions about Ford car models, it became challenging to prevent the retrieval of numerous irrelevant memes related to the actor Harrison Ford, as well as the words “fiesta”, “escort”, or “focus”, which also correspond to certain Ford car models. Therefore, before exploiting a collection of social media posts, it is necessary to perform some additional quality management operations to assess its overall quality and to filter the posts that are relevant for the specific analysis task (Tilly et al. 2017).

Previous frameworks for SoBI and SMA have not paid the required attention to data quality, and more research is needed (Alrubaian et al. 2019). As noted by Stieglitz et al. (2018), in the papers that already document the data tracking and preparation steps of their social media analysis projects, these steps are often dealt with superficially and never with as much extension as data analysis tasks. The authors conclude that the phases of data discovery, collection and preparation of social media data projects require more research. Most approaches just apply a series of ad-hoc rules to posts (e.g., tweets with more than three retweets, users with more than one hundred followers, and so on) to filter out those to be analyzed (Choi, J., et al.

2020; Arolfo et al. 2022). Authors do not explain how the concrete filtering rules have been identified nor how to measure their efficiency to increase the quality of the collection. In general, in the task of building collections of social media posts for analytical applications, two important quality management issues are missing: general purpose data quality models and systematic methods for identifying the best quality metrics for the posts.

1.1 Our Approach and Contributions

To clarify our approach, we must first define the key concepts of our quality model. A *quality attribute* is a qualitative property of the data that expresses some aspect to improve from the analyst's perspective. As we will see, credibility is the most frequent quality attribute for social media data, followed by trustworthiness, reliability, credibility, veracity, relevance, and validity, among others. The concept of *quality metric* refers to any method or function that serves to estimate the level of achievement of a quality attribute for a collection of data. These metrics are typically quantitative, generating numerical values that we denote as *quality measures*. These measures may be presented to analysts in different formats, or they contribute to the calculation of *quality indicators* that combine several metrics to implement more complex metrics.

Selecting the best quality criteria for social media data is a complex task that requires a deep understanding of both the context and objectives of analysis. In the literature, data provenance is the main quality dimension considered for social network analytics applications (i.e., the credibility of the author of the post), yet it has been superficially treated with ad-hoc combinations of aggregated metrics such as the number of likes, mentions or followers, and without considering any contextual circumstances. However, the credibility of a post depends largely on its intrinsic properties and the role the poster plays in the subject of analysis. More specifically, we believe that the credibility of most social media users can be well understood by measuring several aspects in their account definitions, in their metadata and profile descriptions.

In Berlanga et al. (2019), we presented a method to build indicators to assess the overall quality of collections of social media data by integrating the measures obtained by several quality criteria. By considering the peculiarities of each SoBI project (e.g., its context, objectives, topics, and participants), this method helps to find the quality criteria that best suit both the participants and the available posts data, and then integrate them to form a valid quality indicator. This approach relies on the selection of a ranking of relevant users associated with the different categories of posters and taking this ranking as reference, the method automatically calculates the impact of each quality metric. This method was included as a complementary component of the SoBI workflows in Aramburu (2021).

In this paper, we propose a new integrated approach for data quality assessment in SoBI projects, where quality is assessed at the same time that analytical facts are extracted from social media data. In this paper, we extend our previous work by providing a new formal framework that allows the definition of quality indicators adapted to the specific analysis tasks of a SoBI project. More specifically, our approach contributes to the current state of the art in the following aspects:

1. It provides a novel and formal method to measure data quality in social media data according to the main aspects identified in the literature, namely: credibility, reputation, usefulness and completeness. Our method relies on metadata-based metrics as well as content-based metrics derived from language models.
2. Our approach defines a novel multidimensional model, Q-cubes, to capture and profile the quality metrics. This model's main feature is the dimension of user roles, which allow a better understanding of the data quality with respect to their authors. The multidimensional model covers the three main kinds of factual data handled in SoBI, namely: posted contents, involved users, and mentioned topics.
3. The combination of quality indicators and Q-cubes allows analysts to define in a straightforward way the necessary data filters to obtain high quality collections for analysis. Moreover, Q-cubes provide an overall picture of the main features of the extracted facts in terms of their contents and the users that generate or interact with them.
4. The proposed approach integrates both fact extraction and data quality assessment from the beginning of the process (i.e., data sources) to the final output for analysis. The method provides analysts with global quality indicators, as well as partial quality indicators per topics and/or user roles, together with the necessary quality metrics thresholds for filtering the data.

1.2 Organisation of the Paper

The paper is organized as follows. Section 2 reviews the main related work with respect to quality assessment in SoBI projects. Section 3 describes the main quality aspects considered in this paper. Section 4 presents the proposed approach for quality assessment. Section 5 is devoted to the experiments carried out over two long-term data streams and their results. Finally, Sect. 6 provides conclusions, limitations and presents the future work.

2 Related Work

From the point of view of quality management, building collections of social media data for analysis applications involves several things. On the one hand, it is necessary to identify the best group of topics for retrieving the posts from social networks. Here, the purpose is to obtain a set of posts as homogeneous and complete as possible with respect to the objectives of analysis, without biases or missing information. On the other hand, it is also necessary to perform some cleaning operations to select those from the retrieved posts that are really related to the subject of analysis as well as to validate the overall quality of the final collection. To this end, it is necessary to determine the best quality metrics to be applied by means of filtering operations as well as to assess the quality of the overall collection. In both cases, the goals and circumstances of the analysis operation at hand determine a context that is of primary importance when defining these quality operations (Arolfo et al. 2022).

In this section, we review previous methodologies for the construction of social media data collections from the point of view of quality management. Next, we summarize the different approaches to measuring credibility, which is the most frequent quality attribute for social media data. Finally, we review and categorize the remaining quality attributes for social media data.

2.1 Previous Approaches

As explained in the introduction, most papers on social media data analysis work with ad hoc constructed collections. Although many of them recognize the importance of data cleaning operations, they do not explain how to ensure data quality during the preparation of a data collection for real-world scenarios. The framework for enterprise social media analytics proposed by Holsapple et al. (2018) is complete in all aspects of SoBI and SMA systems, but its processing method does not consider any data cleansing and quality assessment tasks. Next, we review a set of papers that have been selected because they illustrate the different ways in which previous approaches to the construction of social media data collections for SoBI and SMA applications have incorporated data quality management operations into their methodologies. Briefly, these approaches range from such that do not provide sufficient support to data quality, via those based on manual or black box methods, through to those focused on social media data streams, and finally up to recent approaches that allow for the definition of some parameters in order to adapt their behavior to the application context.

The uniform data management approach of Goonetilleke et al. (2014) reviews three main groups of research challenges to address when building a Twitter data analytics platform. For data collection, the main issue is the specification of the best set of retrieval keywords and hashtags. For data pre-processing, they demand specific text processing and information extraction strategies for Twitter data. Finally, for data management, they explain that quality management is a major issue, and quality metrics such as trust in authority or authenticity should be included in user languages to query social networks. Consequently, although this paper identifies all these limitations of the available technology, it does not provide a methodology to address them in the proposed framework.

The methodology for SoBI of Abu-Salih et al. (2015) proposes to execute cleaning operations to remove dirty data and ensure data consistency at the data acquisition stage prior to data storage. Later on, during data analysis, the collected data is processed to infer a domain-based value of trust for the relevant data based on the credibility of the data producers. Trustworthiness is estimated by means of a set of key credibility metrics (i.e., number of likes, retweets, replies, …) whose measures feed various machine learning modules proposed to predict high-influential users in a domain (Abu-Salih et al. 2020). In this way, the exploited social media data acquires a minimum level of trust with respect to its domain. This methodology does not include any tools to assist with the selection of the best quality criteria for cleaning operations nor any credibility measures for trustworthiness estimation.

A second methodology for SoBI (Francia et al. 2016) recognizes that crawling design can be one of the most complex and time-consuming activities and aims at retrieving in-topic clips by filtering off-topic clips. They also explain that filtering off-topic clips at crawling time could be difficult due to the limitations of the crawling languages and propose to filter them at a later stage by using the search features of a document's database. The authors note that manually labelling a sample of the retrieved clips enables the team to trigger a new iteration where the crawling queries are redefined to remove off-topic clips more effectively. However, this work does not consider the quality of data as a main objective, and it does not deal with the question how to obtain a good set of quality measures.

In the quality management architecture for social media data presented in (Pääkkönen and Jokitulppo 2017), the data acquisition, data processing & analysis, and decision-making phases can include functionalities for quality control and monitoring. In this approach, data quality management consists of assigning values to a predefined set of quality attributes that depend on the purpose of the data set at hand. In the following, data quality can be evaluated from the point of view of the data source (i.e., data provenance), the data (i.e., data quality) and the user (i.e.,

trustworthiness). The quality, organizational and decision-making policies of the organization define the criteria to filter the quality data. Although the proposed architecture can represent all these data quality elements, the authors do not propose a methodology for defining and applying them.

Over the last few years, several software architectures have emerged to process social media posts in near real-time for analytical purposes. For example, the work in Hammou et al. (2020) is a distributed intelligent system for real-time social big data analytics. This system takes advantage of distributed machine learning and deep learning techniques for enhancing decision-making processes. After data ingestion and storage, and before the text embedding translations, some cleaning operations can be executed. However, these operations only serve to perform a series of pre-processing actions such as removing the numbers, URLs, and hashtags.

Alternatively, Podhorany (2021) proposes an advanced architecture and workflow based on Apache Hadoop and Apache Spark Big Data platforms for collecting, storing, processing, and analyzing intensive data from social media streams. It uses text analysis methods and location estimation techniques to analyze the reported situation by using the information included in the processed posts. Although during the experiments, a cleaning phase executes various filters and text adjustment techniques, data cleaning operations were not included in the architecture proposed in the paper.

Finally, in a recent work, Arolfo et al. (2022) demonstrate the reliability of Twitter data for decision-making processes by means of a software tool that processes streams of tweets for presenting several graphics with quality measures. Its quality model considers four dimensions (i.e., the reliability, completeness, usefulness, and trustworthiness quality attributes) and measures them via a set of basic metrics whose measures are available in the tweets. The user can dynamically adjust the weights of the four dimensions to fit different contexts or interests. This approach constitutes a first attempt to define a context-aware quality model for social media data, but it is still quite limited because it relies on a fixed and non-validated set of metrics. For example, it measures usefulness in terms of sentiment expressions, which is only valid for a concrete type of applications (i.e., sentiment analysis). Similarly, measuring trustworthiness according to being a verified user or having many followers is also a way of restricting the interpretation of this quality dimension. In general, a context-aware data quality model should, first, allow users to define their own application-specific set of metrics to measure quality attributes and, second, provide them with formal tools to validate and choose the best metrics for each concrete analysis task.

2.2 Credibility and Reputation Metrics for Social Media Data

Credibility is the most frequent quality attribute for social media, and many different approaches have been proposed to measure it (Viviani and Pasi 2017; Alrubaian et al. 2019). The literature review clearly reveals that many SMA projects aim at analyzing concrete events such as a catastrophe or a terrorist attack where the main issue is to evaluate posts' credibility (Gupta et al. 2014; Kaufhold and Christian 2020; Saroj and Pal 2022). Customer review analytics is another large application field of social media data and also here credibility is the main issue (Hu et al. 2020; Zheng 2021). It is important to clarify that, for all these works, credibility is a broad concept that intersects with other semantically related quality attributes such as trust, reliability, believability, veracity, relevance, validity and, in some cases, even understandability and reputation.

Among the numerous metrics that feed into these algorithms, some are derived from processing post content, primarily focusing on textual attributes, writing styles, linguistic expressions, sentiments, and additional elements such as URLs or images. A second set of metrics is based on social parameters extracted from post metadata, including information about each post and its author. Lastly, there is a category of metrics that provides insights into the behavior and actions of users within the social network. Table 1 shows a sample of state-of the-art metrics used to measure credibility (Sikdar et al. 2013; Gupta et al. 2014; Viviani and Pasi 2017; Alrubaian et al. 2019) which, in many cases, could also be applied to assess other quality attributes. The broad spectrum of metrics demonstrates that credibility can be interpreted in diverse ways. It is the responsibility of the user to select the most suitable metrics for each project, considering its domain, available data, and the applied technologies (Aramburu et al. 2021).

The review of Alrubaian et al. (2019) found that most related work on Twitter content credibility assessment was performed at four levels of feature extraction: post, user, topic/event (computed as a numerical score for each tweet regarding that topic/event), and hybrid levels. Most approaches use automated and semi-automated techniques, including supervised and unsupervised machine learning algorithms, weighted algorithms, and graph-based methods. Data-driven models classify social media data as credible and not credible, which makes their results difficult to understand for users as they do not receive feedback on the quality features of credible posts. Alternative approaches based on various criteria are emerging, which focus on aggregation schemes to assess an overall credibility estimate (Pasi et al. 2019). Finally, graph-based approaches exploiting the social structure of connected

Table 1 Sample of metrics to measure credibility in social media data found in the literature (Aramburu et al. 2021)

Posts Contents	Posts and Posters Metadata	Users Behavior
# Chars/words	Account age	# Retweets
# Punctuation symbols	Listed count	# Tweets
# Pronouns	Status count	# Tweets favorited
# Swear words	Favorites count	# Mentions
# Uppercases	# Friends	# Tweets are a reply/retweet
# Emoticons	# Followers	Mean time between tweets
#URLs/images	# Followings	# Likes received
# Hashtags	Ratio of followers to friends	# Directed tweets
# Misspelled words	Mean text length in tweets	# Users that propagate the user
# Sentences	Mean hashtags in tweets	# Users the user propagates from
Average length of sentences	Mean # URLs/ mentions in tweets	# Tweets propagated by other users
# Product mentions	Verified account	# Users that converse with the user
# Product features mentioned	User image in user profile	Mean number of conversations
# Opinion sentences	Tweet geographical coordinates	Average length of chain-like behavior

entities analyze credibility propagation in social networks (Viviani and Pasi 2017).

The work of Pasi et al. (2019) proposes a multi-criteria decision-making approach aimed at assessing the credibility of user-generated contents. It considers features connected to the contents, the information sources and the relationships established in social media platforms. Then, the users are asked to manually evaluate all these features in terms of their impact on veracity. By considering different aggregation schema for the partial performance scores and their impact, the authors calculate an overall score of veracity. With respect to data-driven approaches based on machine learning techniques (Crawford et al. 2015), their approach enhances user awareness of the data features influencing the proposed decision, thereby reducing the problem's reliance on specific data. Furthermore, they also argue that making a binary decision on the credibility of a tweet is difficult in most contexts, and it would be better to provide users with both a binary classification and a ranking of credibility.

Finally, Abu-Salih et al. (2019) consider that adding a user-domain dimension to credibility assessment enhances understanding users' interest, but the literature shows a lack of approaches for measuring user-based trust. In particular, the accurate classification of the users' interest assists in providing a better understanding of posts contents. Previous work frequently considers simple measures such as the number of followers to calculate indicators of users' credibility, i.e., when users are in many Twitter lists and have many followers it is because the contents they generate satisfy many users. These approaches ignore that in a domain, users' interests can be diverse and evolve and change over time. To account for this, Abu-Salih et al. (2020) propose to consider this quality attribute as a time and domain-dependent parameter.

Regarding reputation, most work has been focused on identifying the influential users in a specific domain (Amigó et al. 2014). Existing approaches mainly rely on metrics similar to those presented in Table 1, plus combinations of the "followers" and "friends" metrics (Cresci et al. 2015) and vocabulary-based signals (Rodríguez-Vidal et al. 2019). More recent works showed that influential users can be effectively identified by their language models (Nebot et al. 2018; Rodríguez-Vidal et al. 2019).

2.3 Quality Attributes of Social Media Posts for SoBI Applications

While we have previously discussed how credibility can be assessed by combining user and post metrics, we note that the remaining quality attributes can be defined based on post content, user characteristics, and topic dimensions. Below, we review the most significant attributes according to this classification.

The work of Salvatore et al. (2021) also defines a set of quality dimensions and indicators for Twitter data, building upon the framework proposed by Cai and Zhu (2015) for Big Data. In this work, quality was represented into five dimensions: availability, usability, reliability, relevance, and presentation. Authors noted that quality categories are not independent of each other, as changes in a quality dimension impact other dimensions as well, for example, improving data completeness may lead to a loss of data accuracy. The resulting framework was oriented towards the identification of the main sources of error by means of a set of indicators and a collection of good practices that

should be undertaken when using social media data. Although this work is far from being a method to help to find the best quality criteria for a particular social media data collection and analysis task, the complete list of quality attributes that it contemplates are also considered in the following review.

2.3.1 Post Contents Quality Attributes

Two important attributes for measuring post quality are the legibility and clarity of post contents. Previous research has assessed these attributes by using metrics like the readability features proposed by Duan et al. (2012), as well as sets of facets related to linguistic quality from Berardi et al. (2011) and Gupta et al. (2014). The main purpose of these approaches was to discard the posts that were difficult to understand or not very credible because of their linguistic deficiencies. However, to filter quality posts and represent them in a format that is easy for analysis applications to process, the preparation phase should go beyond the linguistic properties of the texts and should try to extract their meaning, which would help users to recognize the semantic elements that are useful for analysis tasks (Kolajo et al. 2020).

Accuracy is an important quality attribute that ensures that the data is free of error. In the case of social media data, this attribute is difficult to measure due to the lack of a comparison baseline (Shankaranarayanan and Blake 2017). In the case of social media posts, the accuracy can be analyzed at two levels: posts contents and user metadata and are in both cases very difficult to measure.

Another important dimension of data quality is timeliness. Among the quality attributes of a domain, it may be useful to define the period during which the posts will add value. These time properties will depend on the objectives and circumstances of the analysis tasks. For example, a review of a car model could be valid for much longer than its promotion at a fare, since it could last for many months, until the manufacturer launches a new edition of the model or it disappears from the market.

The value of posts lies in the usefulness of the data contained in them, in the sense that it should be possible to extract from their contents the values that analysis tasks require (Berkani et al. 2019). Here lies an important source of risk which is the availability of metadata. For example, some analysis tasks, such as segmenting the market opinions with gender, age, location, or profession attributes, require metadata. Social media users do not always provide their real profiles so the available metadata may lack key attributes for the analysis. In some cases, it can be helpful to infer some of this data by semantically processing the content of all the posts, although this is difficult to keep updated for every user (Hernandez et al. 2013).

2.3.2 Users Quality Attributes

Social media users are followers of other users, so considering them as a source of business information makes the author's reputation a quality attribute of utmost importance. Valuable posts come from users with good reputation, because this fact conveys the credibility and accuracy of the contents that they post. The literature frequently considers the number of followers, likes and retweets as indicators of good reputation, e.g., the users that appear in many Twitter lists and have many followers generate posts that satisfy many users. However, reputation is a quality attribute that depends greatly on the business domain. Therefore, measuring the quality of posts cannot be as easy as checking the number of followers of their posters, it also requires considering further domain dependent conditions.

In social media platforms, a user account is verified if it proves to be a public interest account. Users with professional purposes will obtain better results by using verified accounts. Verification standards are clear and, among other strict conditions, the user account definition must contain serious information including a profile description, header photos, name, biography, and location. In general, account profile descriptions delimit the role of the relevant users in a business domain or application context. Therefore, the quality of verified users' accounts should always be considered together with their profile descriptions.

2.3.3 Quality Attributes for Topics

In social media platforms, there are mechanisms to retrieve posts by means of keywords, usernames, and hashtags (Goonetilleke et al. 2014). However, data completeness is not ensured due to the following causes:

- Using keywords there is no certainty of retrieving all the posts that deal with the subject, therefore bias and data loss may occur (Plachouras et al. 2013).
- It is difficult to find the set of hashtags that must be part of an analysis subject (Bansal et al. 2015).
- It is almost impossible to identify all the representative users of a topic of analysis, and a percentage of representative voices will be lost due to the simple fact that they have not participated in social media (Czernek 2018).

Rather than completeness, topic coverage is a quality attribute for social media that indicates whether the query used to retrieve a collection's posts is complete in the sense that it includes all relevant topics related to the objectives of the analysis task, considering keywords, hashtags and usernames as the representative elements of a topic of analysis.

2.4 Main Conclusion and Methodology of Work

In this section, we have reviewed relevant methodologies to build collections of social media data from the point of view of quality management. The main conclusion is that while most approaches to social media analysis for decision support apply different quality criteria during data preparation, it is not clear at this point how to define a general-purpose quality management method. Previous work has proposed many different quality metrics for multiple purposes, which depend mainly on the respective application, but whose effectiveness and validity is unproven. The experience demonstrates that, whatever method applied to assess the quality of data, a good combination of different types of metrics is part of the solution. However, there is no systematic methodology for identifying a valid set of quality metrics to build a reliable collection of social data for analysis tasks in a domain of application.

Our work methodology can be classified as design science research (DSR) (Johannesson and Perjons 2014). Initially, we identify the overarching issue of data quality in social network data, a topic extensively studied in the literature. Furthermore, we observe the absence of a well-founded methodology for assessing data quality in social network analysis. Consequently, in the subsequent sections, we propose the utilization of two grounded theories to address the data quality problem: multidimensional data modeling, and information retrieval (IR). The former explains the collection and summation of metrics to form quality indicators, considering various perspectives of the data quality issue. The latter deals with the relevance ranking of quality metrics. Our primary hypothesis starts from the assumption that data quality is significantly influenced by both the relevance of its posters for and the coherence of their posts in relation to the application domain. As a result, the solution development is consistently supported by the chosen theories and premises. Finally, the proposed method provides essential information to measure the quality of analytical data and make decisions about data filtering and/or parameter updating. Consequently, the evaluation of the resulting dataset by analysts may imply redefining some of the parameters of the entire extraction process, such as the keywords used to retrieve the data and the set of reference users. Subsequent iterations can then be carried out to further enhance the dataset. These iterations should always be guided by the automatically derived quality indicators, which demonstrate whether the actions taken have improved the results.

3 Data Quality Management Dimensions

Nowadays, data quality management is considered to be one of the main factors that guarantee a successful adoption of AI technologies by modern business and organizations (Jöhnk et al. 2021). As explained in Sadiq and Indulska (2017) and Zhang et al. (2019), traditional methods for managing data quality follow a top-down user-centric approach: the analysist specifies some quality rules that serve to govern data, to assess data quality, and to execute cleaning operations. This approach is suitable for managing the quality of data generated internally by an organization. However, when the organization does not control the external processes that generate the available data, as in the case of social media, quality assessment requires prior knowledge about the data features. To gain this knowledge, data quality management follows a bottom-up approach that starts with submitting the source data to some exploratory tasks (Zhang et al. 2019). These tasks help to find data quality rules and requirements that will drive the data collection process. To execute the preliminary exploration of the available data, interactive, statistical and data mining techniques are applied (Stieglitz et al. 2018).

Social media posts present many distinct aspects that could serve to filter them, with posts contents and users' attributes and interactions being the main contributors to quality metrics (see Table 1). However, the selection of the best quality metrics for a specific SoBI project requires a deep understanding of its business context, strategy, and objectives of analysis, as well as of the relevant social media data (i.e., posts and users) to be managed (Immonen et al. 2015; Berlanga et al. 2019). Thus, we consider three different dimensions for data quality: the social media users, the posts they generate, and the topics they write about. As Table 1 shows, these dimensions have been widely adopted in most of the approaches of social media analysis. They provide different quality metrics whose convenience, in the case of users and posts, will depend on the types of users that participate in the business domain.

In our work, we propose performing *global quality analyses* over long-term data streams. This is because quality problems, such as redundancy, bias and noise are often difficult to detect by means of local analysis (i.e., directly over the streamed data). The other strategy we propose for data quality management is *profiling the long-term data stream* according to a series of quality dimensions. Basically, as we will explain in following sections, profiling is performed by analyzing the language models of the users' profiles and their posts according to the intended quality analysis dimensions.

3.1 User Dimensions for Data Quality

Social media data profiling allows the analyst to have a better understanding of the real market of the business domain. For example, finding the most frequent topics in a collection of car reviews can help us to identify the range of aspects that should be part of the product features analysis dimension, as they are the hot topics in the market. Furthermore, profiling the range of users that post on the domain along with their metadata is also important for determining the measures and dimension attributes available to take part of the analysis multidimensional data structures (i.e., cubes). For example, demographic data in user descriptions can help defining the attributes of the customers' analysis dimension. Similarly, classifying the range of users who post about the car models of a brand into the different stakeholder groups can help the analysis' purposes in many different ways.

One main novelty of the proposed model is that we profile social media users according to business-related classes. For example, most verified user accounts have a strong relation to professional purposes, and their definitions contain useful information including a profile description, header photos, name, bio, and location. In this paper, we profile the users of a generic business domain according to the following main categories:

- **"Domain Business Users",** which have on-domain professional/business profiles and apply social media accounts to promote their products and services by posting high-quality contents regularly. They are often verified users.
- **"Domain Influential Users",** which can be identified by their profile descriptions and their large number of followers and retweets. In case they are unverified users, other users give them authority, and as experts they often publish quality posts for that domain. Influential users are followers of business users.
- **"Domain Interested Users",** which are relevant because of the high level of similarity between the domain and their profile description. Usually, interested users are followers of business and influential users.

Figure 1 shows some examples of this classification applied to the automotive domain. This classification allows analysts to distinguish users with clear roles from those whose relationship is more sporadic or irrelevant. In general, the credibility of the users with a clear role in a business domain and the quality of their posts is higher than that of the rest of out-of-business users.

3.2 Social Media Data Quality Perspectives

In this work, we define four perspectives for social media data quality, namely: credibility, reputation, usefulness, and completeness. These perspectives are derived from the discussion presented in Sect. 2. They serve to classify the chosen quality metrics and facilitate their combination into specific quality indicators to estimate the degree of achievement of each quality perspective.

Credibility indicators must reflect how reliable the user accounts are. Reliability means that the users are real and relevant to the analytical goals, and that they post information that can be trusted when performing an analysis of these data. Measures related to credibility are primarily associated with the activity of the users, the coherence of the contents they generate, and other evidence that characterize good posters. Users whose intentions differ significantly from the expected ones should be assigned an extremely low value for credibility. For example, spammers and jokers should be categorized as of low credibility.

Reputation indicators should consider the factors that contribute to user influence, and, therefore, the impact of the content they generate. Usually, high quality is associated with reputed accounts. However, in some domains, highly influential users are not aligned with the analytical goals, being the contents, and generate useless posts for the analytical goals at hand. In this case, although it is always desirable to have a suitable number of reputed accounts, there must be a trade-off with respect to other quality perspectives, such as usefulness.

Usefulness indicators are of primary importance as they give us the clues of the potential impact of data on the analytical tasks. These indicators mainly measure the relevance and readability of the extracted data to derive useful facts for analysis. In this paper, we introduce the concept of coherence, which aims at measuring how well the languages of the data stream and the analytical goals are aligned. We will define the usefulness indicators over these kinds of measures.

Lastly, **completeness** should be viewed as a measure of how well the data covers a specific analytical topic. In this case, data has already been transformed into facts and we can directly measure how well facts cover the desired dimensions of analysis.

In the following section, we propose a new multidimensional model that integrates the elements defined in this section (i.e., user categories and quality perspectives) as a way to improve the analysis of social media data quality from the point of view of the different types of users participating in the application domain.

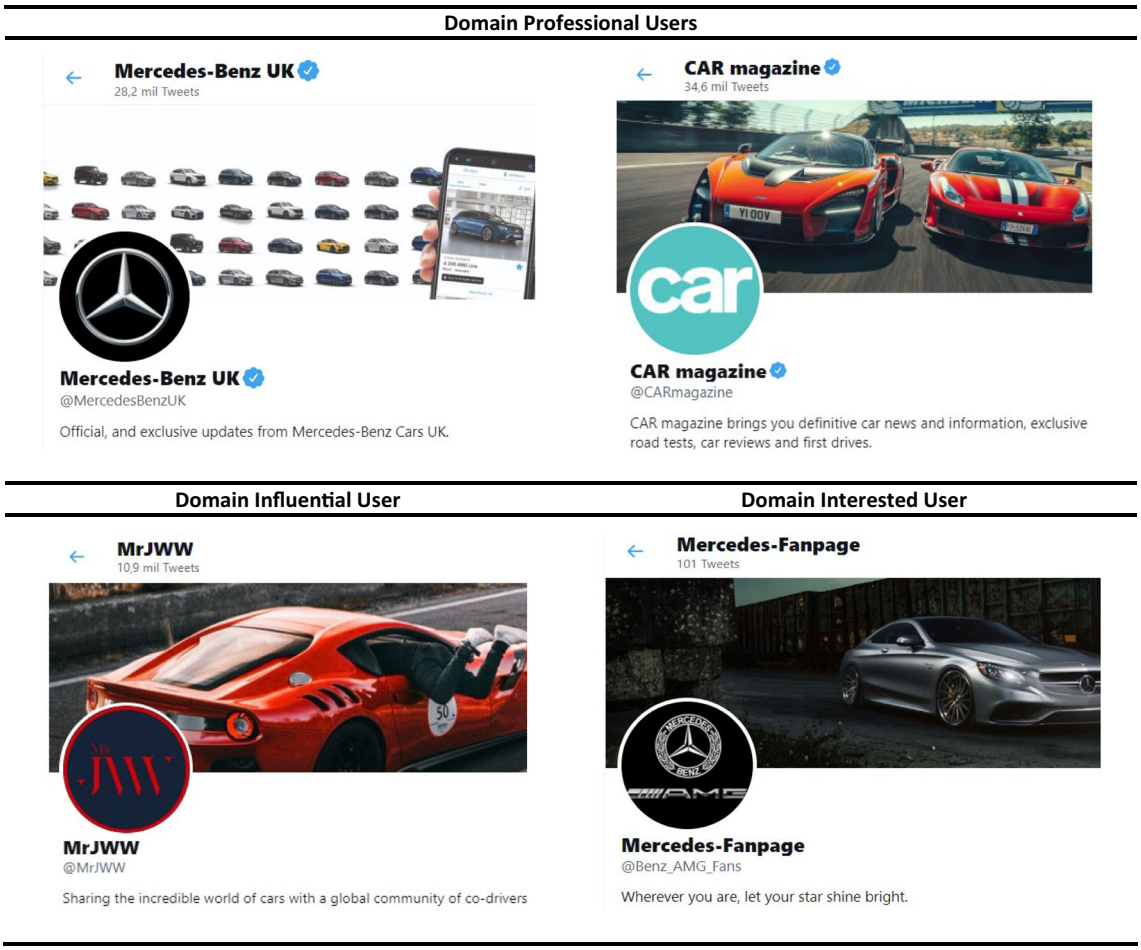


Fig. 1 Examples of Twitter accounts in the automotive domain for each category of relevant user

4 A Multidimensional Model for Quality Assessment

Following the multidimensional data model for Business Intelligence systems (Kimball and Ross 2013), we define several data structures (i.e., cubes) for representing the different quality perspectives as required by analysts. In these cubes, we store measures of the quality metrics' impact to filter out high quality posts for each type of user. Thus, we define two separate dimensions that represent the different post quality metrics and the user roles that can help to decide on the appropriate quality criteria for each analysis task. In this section, we develop the elements of this quality model whose main purpose it is to assess its metrics for a domain-related data stream.

4.1 Facts Extraction and Analysis

The process of social media analysis starts with the definition of a posts data stream using the social network API. The data stream is configured with a series of keywords that are directly related to the goals of analysis. Quality analysis can serve users as a guide to assess the effectiveness of the chosen keywords and the potential lack of data for the intended analysis goals. Figure 2 summarizes the process of social media fact extraction and the subsequent process for quality assessment, which is described in turn.

At this point, it is important to note that Fig. 2 consists of two parts. The lower part (shaded in grey) corresponds to the extraction of facts from the data sources and is not treated in this paper because it is part of our previous work on the SLOD-BI infrastructure (Berlanga et al. 2015). The upper part of the figure includes the Aggregation and Quality Assessment phases and constitutes the central contribution of this work, namely, a new data processing method for quality assessment for social media analysis. In the following paragraphs, we will briefly explain the main components of Fig. 2.

Analysts design their goals by choosing the topics of interest and associating them to a series of analysis dimensions and measures. For example, the topic "car recalls" will have associated dimensions like "location",

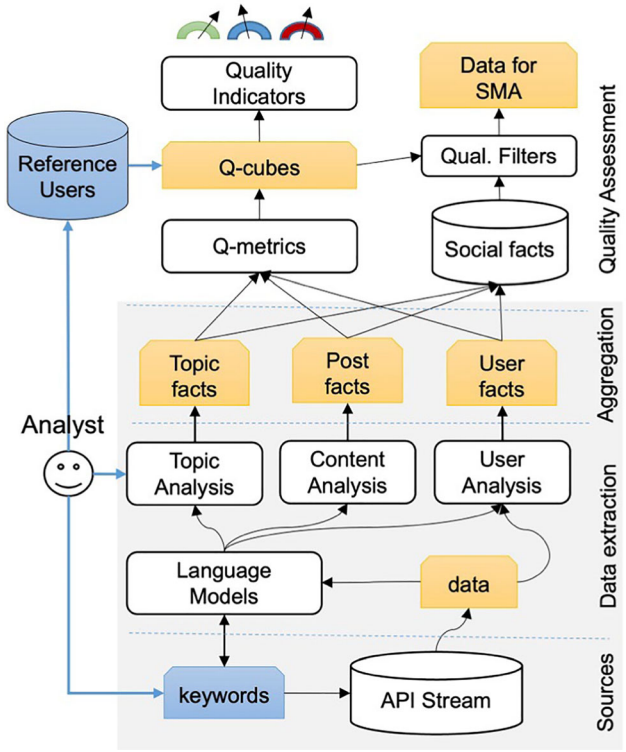


Fig. 2 Proposed data processing method for quality assessment for social media analysis

"detected failure" and "car model" and measures like "reported_cases" and "social media impact". Thus, we assume that the analyst has defined the set of dimensions D that are of interest for their analysis. For the sake of simplicity, we define an *analysis dimension* $d_i \in D$ as a set of values where each value is associated with a description and the lexical elements that enable recognizing the value in the post's contents and metadata.

Once data are retrieved from the social network API, we process them from three different perspectives: topic analysis, content analysis and user analysis.

Topic analysis concerns the discovering and organization of themes of interest from a large collection of posts. Topic analysis is one of the central tasks of any social media analysis as it serves to gain insight into the main concerns of social media users. Its main challenges are the high dynamicity and semantic drift of the user-generated contents, which make it necessary to continuously track the stream. Unsupervised machine learning has usually been adopted for topic analysis, mainly clustering and statistical methods like n-gram analysis and Latent Dirichlet Allocation (Chauhan and Shah 2021).

Content analysis is primarily focused on extracting implicit information from user-generated contents such as sentiments (i.e., polarity), emotions and entity mentions relevant to the analysis goals. Both supervised and unsupervised machine learning methods have been proposed in the literature for this purpose (Birjali et al. 2021). The output of the content analysis module usually are the facts that end-users are supposed to analyze. Content analysis is directly guided by the analysis dimensions and measures defined by the analyst. The facts extracted through content analysis will be inserted into the fact tables for performing the integrated BI tasks.

The User analysis component assigns a profile to each user account according to the analytical goals. Author profiling is a related task that aims at identifying user attributes from their generated content and biographies. Approaches in the literature have mainly focused on demographic attributes such as age, race, and gender. Some works have also treated more interesting attributes for BI such as professional profiles and influence degree (Amigó et al. 2014; Han et al. 2017; Nebot et al. 2018; Rodríguez-Vidal et al. 2019).

These components produce three types of facts as output: topic, post, and user facts. These facts represent all the explicit and implicit data that are useful for analysis. Therefore, our aim is to measure the quality of these facts and propose methods to improve their quality for analysis tasks.

In this paper, we assume that these components are dealing with a collection of posts C, from which a series of facts are extracted, denoted as *facts*(C), which can be further distinguished to be topic facts (*t-facts*), user facts (*u-facts*) and post facts (*p-facts*) when necessary. Finally, we can filter the extracted facts by applying the quality criteria derived from the quality cubes. In the following sections, we discuss how to measure the quality of social media data in terms of these facts and the set of *Reference Users* whose posts are recognized to possess of good quality.

4.2 Q-cubes: Multidimensional Analysis of Quality Metrics

Our approach defines a novel multidimensional model, consisting of three quality cubes (Q-cubes), to capture and profile quality metrics. Specifically, we propose two Q-cubes for analyzing the quality metrics of a domain-related data stream, namely: the Posts Quality Cube (PQC) and the Users Quality Cube (UQC). In addition, to assess the quality of the posts for each specific analysis topic, we also define a third cube called the Topic Quality Cube (TQC).

The PQC aims at measuring the impact of quality metrics derived from the contents and metadata of the posts. Table 2 summarizes the main aspects regarded for the PQC cube. Similarly, the UQC aims at measuring the impact of quality metrics associated to different aspects

Table 2 Metrics categories for the Posts Quality Cube

Group Id	Post Attributes	Quality Metrics
P1	Metadata	Metrics provided along with the posts (e.g., #retweets, #likes, etc.)
P2	Contents	Metrics derived from the contents of the posts, like text and images
P3	References	Metrics involving the quality of links, mentions and hashtags included in the posts

Table 3 Metrics categories for the Users Quality Cube

Group Id	User Attributes	Quality Metrics
U1	User's Posts	Aggregate metrics over the posts written by the user (e.g., #tweets on-domain, stylistic-related metrics)
U2	Description	Metrics derived from the description in the users' profile accounts
U3	Metadata	Aggregate metrics of the posters (e.g., #followers, #friends, etc.)
U4	Interactions	Metrics derived from the interactions of the posters and towards the posters (e.g., #performed actions, #received actions, etc.)

related to the users. Table 3 summarizes the main aspects of metrics included in the UQC.

The TQC cube will provide a summary of the quality aspects associated with each analysis topic, as well as the necessary information for selecting the suitable quality criteria for data filtering. Table 4 shows the two groups of metrics we consider for topics. In addition to these metrics, as topics are subsets of posts, all quality metrics in Table 2 can be also applied to topics by aggregating them accordingly.

The Q-cubes are built with the impact values derived from processing the long-term data stream. We use Q-facts tables to store the extracted facts that will serve to fill the Q-cubes, that is, each Q-fact table includes all the observations of the quality metrics for each post/user/topic of the long-term data stream. More details about how PQC and UQC facts are processed and aggregated in stream can be found in Lanza-Cruz et al. (2018).

The three Q-cubes share the User-Role dimension. This dimension regards the classes of reference users for the domain at hand (see Sect. 3), and it is a clear indicator of the credibility of the social media users. Therefore, the role of users becomes the main dimension for assessing the quality metrics. As Fig. 3 shows, the User-Role dimension includes the corresponding Domain Business Users, Domain Influential Users and Domain Interested Users hierarchies. Analysts can further refine these conceptual categories into more specific user profiles for a business domain. For example, in the case study of this paper, we define the following sub-categories: Employees (E), Professionals (P), Public Services (PS), Journalists (J) and Lovers & Fans (L&F). These categories are inspired in the RepLab 2014 dataset and designed according to the experts' criteria involved in the project (Amigó et al. 2014).

The User-Role dimension is populated with a list of Reference Users who are supposed to produce high-quality posts contents. These users of reference must be present in the long-term data stream to compare them with the rest of users. Before we start building the Q-cubes, we need to attach the User-Role dimension to the Q-facts tables. For this purpose, we label as relevant all the facts involving the reference users, and we add the user category labels associated to them. The last step is to build the Q-cubes by measuring the impact of all the included quality metrics. This step is performed as follows:

Table 4 Metrics categories for the Topics Quality Cube

Group Id	User Attributes	Quality Metrics
T1	Audience	Percentage of users deemed relevant to the topic, ratio of users per relevant business roles, etc.
T2	Contents	Number of tweets, contents diversity (e.g., Yule's metric), temporal distribution, coverage of dimensions, etc.

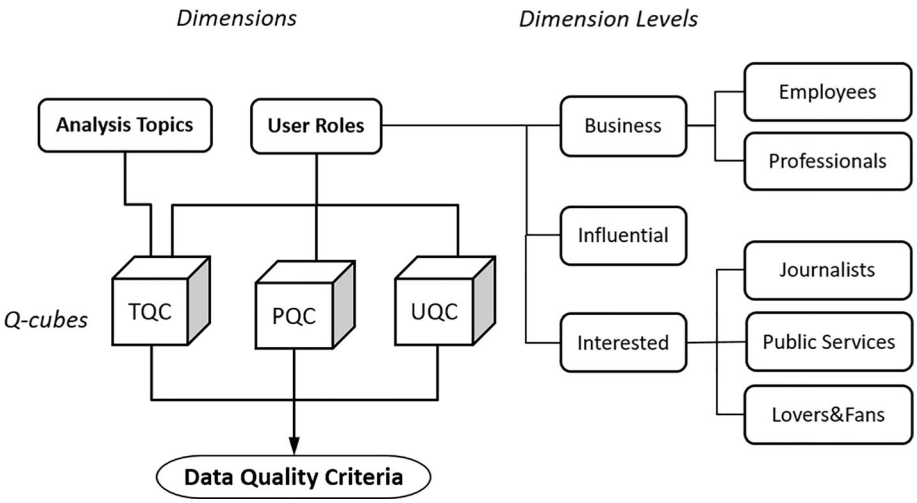


Fig. 3 Quality cubes for social media data

1. For each quality metric, we arrange the Q-fact table based on its values. Being a quality metric, the default ordering should be descendant, that is, the greater the value the higher the quality.
2. We then calculate the overall impact of the quality metric using an evaluation metric applied to the obtained ranking of facts (in the way explained in Sect. 4.3).
3. Finally, we measure the impact of the quality metric for all the categories associated to the User-Role dimension to populate the corresponding cube.

As a result, the final Q-cubes show the impact of each included quality metric broken down by user categories. From these Q-cubes, we can finally derive the quality criteria that allow us to refine the final dataset for analysis purposes.

4.3 Impact of Quality Metrics

To assess the impact of quality metrics, we apply the average precision (AP), which has been widely applied in information retrieval (Baeza-Yates and Ribeiro-Neto 1999). This metric is both easy to implement and efficient to compute. Moreover, recent work has shown how this metric can be approximated with a differentiable function, allowing it to be included in deep learning models (Cakir et al. 2019). Given a list of ordered items (users or posts), the metric AP is defined as follows:

$$AP = \frac{\sum_{k=1}^{N} P_k \cdot rel(k)}{R}$$

where R is the number of relevant items in the collection, N is the size (i.e., number of items) of the complete collection, P_k is the precision at position k, and $rel(k)$ is a binary number indicating whether the element at position k is relevant or not. Notice that if relevant items are uniformly distributed in the ranking, then the value of AP is $AP_{unif} = \frac{R}{N}$

In order to compare impact metrics from different rankings and perspectives, we define a normalized metric that takes into account the relative change with respect to AP_{unif}, namely:

$$AP_{rel} = \frac{(AP - AP_{unif})}{AP}$$

Notice that rankings with AP_{rel} near zero are not useful for quality assessment since they are not able to promote high-quality items. Negative scores indicate poorer quality metrics since their corresponding ranking demotes qualified data.

Figure 4 illustrates the relationship between AP_{rel} and the ranking power of two different metrics. The graphic above represents a good metric, in this case, the number of tweets in the domain, as it promotes reference users to the top positions when ranked by the metric. The graphic below represents a worse metric, in this case, the number of followers, as reference users have been positioned randomly when ranking by the metric.

In this way, we define the impact of a quality metric M, denoted $impact(M)$, as the AP_{rel} measure when ordering the data with M.

It should be noted that the usefulness of a metric is intrinsically defined by how effectively it distinguishes relevant users from irrelevant ones. The analyst might discover some justification for the metric's behavior a posteriori by analyzing its results; for example, the metric "#followers" does not perform well in the automotive domain because many relevant users have a discrete value for this metric. Being familiar with the domain of an application can help analysts identify the set of relevant users, understand the behavior of different quality metrics, and enhance the overall efficiency of the method.

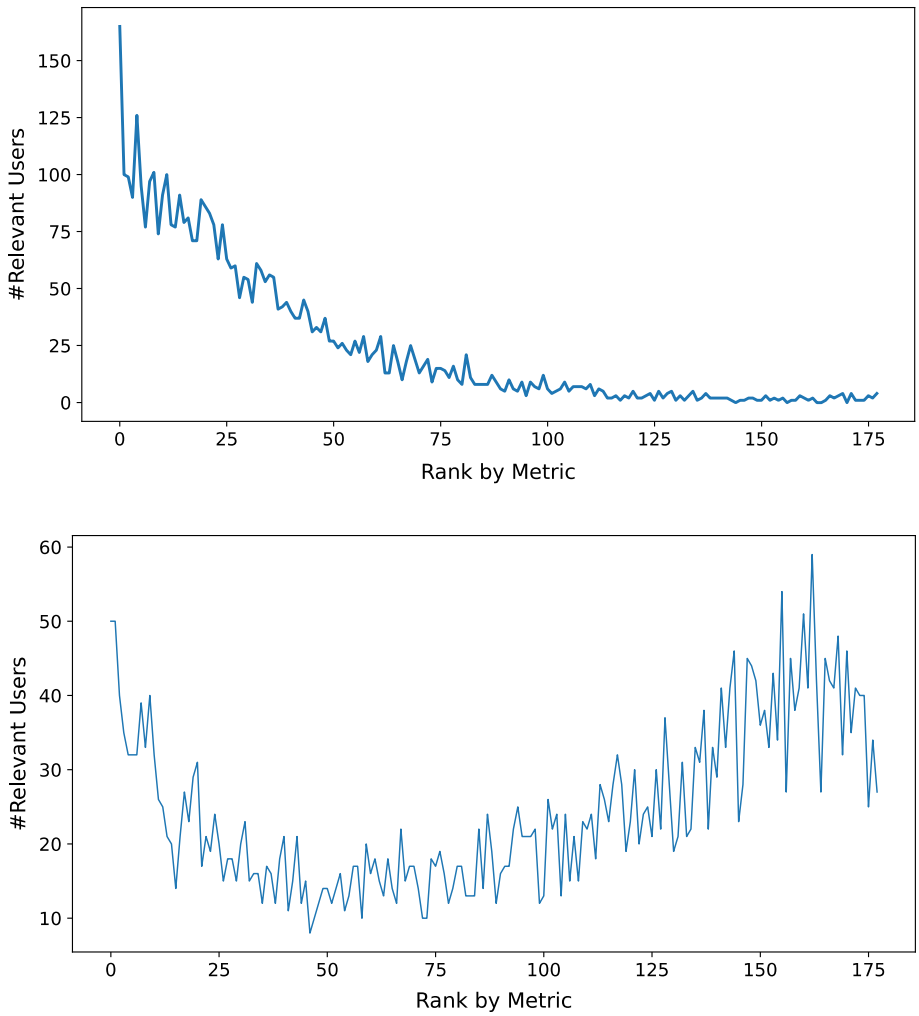


Fig. 4 Examples of the impact of two metrics in the ranking of reference users in the automotive domain (top: #on-domain tweets, bottom: #followers)

4.4 Definition of Quality Indicators

In this section, we explain how Q-Cubes can be applied to define quality indicators for the four quality perspectives adopted in Sect. 3.2. We define a quality indicator as having the following elements:

- A normalized value, usually between 0 and 1, where 1 indicates the maximum quality level and 0 the lowest one; sometimes for convenience we will use percentages.
- A series of quality metrics taken from the available data, which provide the support for the quality indicator.
- A formula to derive the indicator value from the selected quality metrics.

Thus, Q-cubes provide the quality metrics along with their impact for the different analysis facts (i.e., posts, users, and topics). Table 5 summarizes the proposed methods to measure the quality in the different perspectives.

Table 5 Quality perspectives and involved methods

Quality perspective	Methods for quality indicators
Credibility	UQC
Reputation	UQC
Usefulness	Language models + PQC
Completeness	Language models + PQC + TQC

We can define credibility and reputation indicators directly from the Q-cubes metrics, whereas usefulness and completeness depend on the topics and analytical goals at hand. Notice that usefulness measures the relevant facts that we can extract from the data, and completeness depends on the specific dimension values involved in a specific analysis (e.g., mentioned organizations, places, etc.). It is also worth mentioning that completeness depends on how posts are arranged to form the final analytical facts. The common approach is to treat posts

individually (Arolfo et al. 2022), which implies that only a few dimensions can be extracted from each post. To achieve a higher level of completeness, we need to group posts around entities and valid times to capture the different dimensions of the intended analytical facts.

A straightforward method to define quality indicators consists in estimating the percentage of users or posts that fulfil some property (Arolfo et al. 2022). For example, this could be the number of posts mentioning at least one brand name, or the number of posts sent by users with a certain number of followers. However, these properties are difficult to define, and they strongly depend on each specific domain, the analytical goals, and the community of users that generates the contents. In this work, by means of the Q-cubes, we aim at identifying the relevant quality metrics that allow us to distinguish relevant users and contents. As these measures promote reference users at the top positions, we can derive an indicator directly from the distribution induced by them.

We define a quality indicator from a quality metric as follows:

$$\begin{aligned}\text{score}(M) = \ & \text{impact}(M) \\ & \cdot \max_{x \in M}(2 \cdot \text{cov}(x) \cdot q(x)/(\text{cov}(x) \ + q(x)))\end{aligned}$$

This score combines the impact of the metric M according to the corresponding Q-cube, and the maximum of the harmonic mean between the ratio of covered posts and the quantile at a given cut point x in the metric M. In other words, we try to maximize the coverage of posts and the ratio of good posters. A high quantile value combined with a high impact implies high quality, because we are selecting a small number of very relevant users. This value is combined with the covered posts by these users so that we can find a good trade-off between them. The cut point of the metric could be directly used to filter out the data to increase the quality of the dataset. However, this should be only performed when the impact of the metric is high enough.

Credibility and reputation quality dimensions are directly associated with these scores. More specifically, we will take as reference the maximum score of the metrics that are related to these perspectives. For example, metrics like "#followers" are usually associated with reputation, whereas the meta-attribute "verified account" refer to credibility. Notice that these perspectives can take metrics from both posts and users Q-cubes, however, our experiments have demonstrated that user metrics (UQC) obtain much better results.

4.4.1 Language Models for Posts Facts Quality

To account for the usefulness and completeness of the dataset, we propose to use language models. A language model is a probability distribution assigned to each word or term of a vocabulary, which can be further conditioned by a series of contextual parameters (Baeza-Yates and Ribeiro-Neto 1999). As we aim at identifying entities and mentions of dimension values in the dataset, language models are a useful and well-grounded tool for estimating how well the domain is covered by the collection of posts and how well analytical goals are aligned to their contents. It is worth mentioning that previous work on language models for social networks have been shown effective in profiling users by their posts (Nebot et al. 2018; Rodríguez-Vidal et al. 2019).

We define two quality measures based on language models, namely: profile coherence and post coherence. These measures directly calculate the log likelihood of the profile and post contents with respect to a language model L representing the intended analysis goals for a particular domain. Coherence for a fact f is defined as follows:

$$coherence(f, L) = -\sum_{v \in f} \log P(v|L)$$

It is worth mentioning that the lower the metric the higher the coherence. Thus, we must filter out high values to ensure a high coherence. This metric is equivalent to the perplexity of the language model L, which has been also proposed for post quality in (Lin and Morgan 2011).

Usefulness of a post collection C can be defined as the ratio of facts extracted from C that are coherent enough to the language model of intended analytical goals L. This can be formally stated as follows:

$$usefulnes(C, L) \quad = \frac{\#\{f | f \in facts(C) \wedge coherence(f, L) < \delta_U\}}{|facts(C)|}$$

The parameter δ_U can be empirically set from the set of reference users specified for building the Q-cubes.

Completeness can be also expressed in terms of language models as follows. In this case, we need to measure the overlap between the vocabulary and the analysis problem, which consists of a set of dimensions $\{D_i\}$, and the facts extracted from the user-generated contents. The completeness of a fact f is simply defined as:

$$completeness(f, L) \quad = \frac{\sum_{i=1}^{N} \#\{d | \exists v \in f, d \in D_i \wedge P(d|v, L) > \delta_c\}}{\sum_{i=1}^{N} |D_i|}$$

In other words, we measure the ratio of dimension values that can be entailed by the fact values. The entailment relationships between terms are established by using either traditional statistical techniques or via modern word/

sentence neural embeddings (Lauriola et al. 2022). With this proposal, we also account for potential alignments that are not explicitly established between the dimension values and the extracted facts. As an example, the location attribute of a post can be inferred from the user or post metadata. In many cases, these locations are not specified in a systematic way, and it is very unlikely that they match our dimension values for locations. This way to express completeness, will account for hidden associations that otherwise will be lost because of issues like lexical mismatching. The threshold for entailments δ_c can be empirically set up by analyzing historical data.

Completeness can be easily defined for the whole collection just estimating the average over all the facts extracted from the intended collection.

$$completeness(C, L) = \frac{\sum_{f \in facts(C)} completeness(f, L)}{|facts(C)|}$$

Notice that all these measures can be applied to any arbitrary collection of facts extracted from user generated data. Moreover, these facts can be defined at user, post, or topic levels. Thus, for topic analysis we can restrict the collection to the facts relevant for some topic and estimate the previous indicators over the selected facts. In this case, the reference language model L must be also adjusted to the dimensions of interests for that topic.

5 Results

For demonstrating the usefulness of the proposed approach, we have chosen two long-term streams of tweets related to the automotive domain and to natural disasters. The automotive domain stream has been active since 2015 until now and it has served as basis of several studies about SoBI (Berlanga et al. 2015, 2019; Lanza-Cruz et al. 2018). We have generated this stream by specifying a series of keywords related to different car models and brands. It contains around 1,930,617 tweets, written in both Spanish (456,059) and English (1,474,558). The total number of involved users in this stream is 318,469 (up to November 2022).

For the second stream, the theme "impact of natural disasters and migration in the tourism sector" has been defined, which has been created by simply picking up the keywords of the theme, namely: "natural disaster", "migrants" and "tourism" (both in English and Spanish). It is important to notice that these keywords can introduce a lot of noise because Twitter searches for each word individually, not only in posts but also in screen names. In this work, we will deal with the data generated during the period 2019–2022, which contains around 26 million tweets involving around 21 million users. This period includes very popular topics like Brexit and the COVID-19 pandemic.

5.1 Users-Role Dimensions

For the first domain, we make use of two data sources for identifying relevant users. The first one is the RepLab 2014 dataset (Amigó et al. 2014), which contains a track for the automotive domain. We have selected only influencers from this dataset to ensure high-quality users. Thus, we get 480 influential users from RepLab (RL). For the second source, we have analyzed the bigrams language of the user descriptions and we have selected a representative set of bigrams for each user category. Table 6 summarizes the number of selected users along with some bigram examples used to build each category. The total number of reference users from this second source is 28,165.

For the "natural disasters and tourism" data stream, we have mainly focused on three main categories of reference users, namely: people and organizations working in aid/recovery, tourist destinations officers, and journalists. In this domain, the total number of reference users is 7386, from which 1290 are recovery-involved users, 3531 are destination officers, and 2565 are journalists. For this data stream, we have no influential users as reference. Since the automotive domain is much richer in terms of user categories, and for the sake of space, we will show only the quality cubes associated to the automotive domain in the next section. Global quality indicators are shown in Sect. 5.3 for both domains.

5.2 Quality Cubes

Table 7 presents the resulting PQC for the automotive domain. This table only includes quality metrics with a significant impact in some of the user categories. Shadowed cells represent near zero (< 0.1) and negative scores for AP_{rel}, (i.e., they are not relevant metrics for quality assessment). It is worth mentioning that AP_{rel} scores can only be compared within the same column, as the number of references varies by category, resulting in different scales.

In the PQC (Table 7), we have included the metric P1.1 as a fake quality metric so that we can check that effectively it has no impact in the quality assessment. We can see that many quality metrics for tweets have little impact in many user categories. In general, PQC metrics are less relevant than UQC metrics (Table 8). Tweet quality metrics mainly contributed to influencers, especially the metrics favorites (P1.4) and re-tweets (P1.5). Text coherence only contributed to the professional category, indicating that this is the main voice of the stream. The most relevant metric when regarding all users are the use of punctuation

Table 6 Examples of bigrams used for categorizing users in the automotive domain

Employees (E)	Professionals (P)	Public Services (PS)	Journalists (J)	Lovers & Fans (L&F)
2,389	8,705	739	8,611	7,721
·community manager	·used cars	·call emergency	·latest news	·love family
·project manager	·cars service	·crime call	·news information	·sports fanatic
·writer photographer	·car dealership	·report call	·motoring news	·love cars

Table 7 Results for the Post Quality Cube of the automotive-related Twitter stream (in bold face max value per role)

Quality Metrics	User Categories						
	All Categories	Influential Users	Business Users		Interested Users		
			E	P	PS	J	L&F
P1.1 Tweet date							
P1.2 Tweet is reply					0.198		0.311
P1.3 #Tweet replies		0.115	0.316		0.109		0.170
P1.4 #Tweet favorites	0.135	**0.737**	0.414		0.566	0.295	0.192
P1.5 #Re-tweets	0.190	0.716	0.386		**0.858**	**0.339**	0.128
P2.1 Coherence				0.288			
P2.2 #Numeric tokens	0.147			**0.379**			
P2.3 Tweet polarity	0.141		0.197	0.260			0.336
P2.4 Tweet repeats	0.172	0.247	0.124	0.145	0.463	0.215	0.153
P2.5 #Punctuation	**0.205**	0.131		0.157	0.302	0.123	0.318
P2.6 #Emoticons			0.104		0.204		
P2.7 #Mentions		0.122	**0.444**				**0.337**
P3.1 #Links		0.250					
P3.2 Question marks	0.167		0.164	0.233			0.176

Table 8 Results for the User Quality Cube of the automotive Twitter stream (in bold face max value per role)

Quality metrics	All categories	Influential users	User categories				
			Business users		Interested users		
			E	P	PS	J	L&F
U1.1 #Tweets on domain	0.729	0.962	0.834	0.708	0.772	**0.874**	0.624
U2.1 Coherence*	**0.760**	0.663	0.454	**0.859**	**0.780**	0.855	0.604
U2.2 Description length	0.406	0.439	0.595	0.563	0.721	0.350	0.587
U3.1 Account age		0.748	0.139			0.190	0.477
U3.2 #Statuses		0.617					0.118
U3.4 #Followers		0.924	0.367			0.535	0.285
U3.5 #Friends		0.766	0.408			0.367	0.284
U3.6 #Listed count		0.933	0.570		0.594	0.756	0.278
U3.6 Has location		0.329	0.144		0.167	0.158	
U3.7 Verified account		0.651	0.486			0.172	
U4.1 #Performed actions	0.481	0.933	**0.958**	0.254	0.344	0.660	**0.734**
U4.2 #Received actions	0.594	**0.975**	0.867	0.400	0.491	0.857	0.506
U4.3 #Total interactions	0.596	0.974	0.922	0.380	0.463	0.530	0.648

symbols for formatting the message (P2.5), followed by other stylistic-related metrics. Another relevant metric is the repetition of the tweet text: the more a tweet is repeated (not re-tweeted) in the stream the more relevant. Notice that this is valid for the automotive domain where intensive marketing campaigns are frequently performed. In many other domains, this metric would indicate instead low quality because it implies data redundancy.

The most relevant quality metric associated to users (Table 8) is the number of tweets related to the domain (U1.1), which have a high impact in both all users and influencers. The coherence of the description with respect to the domain has also a high impact in quality, being the best metric for the professional category. We can see that the different user categories present different quality metrics profiles, showing thus different behaviors in the data stream. Consequently, the assessment of user quality cannot only rely on fixed quality metrics; instead, it should consider metrics that align with each specific user role.

Regarding the influencers, we observe that most quality metrics have a high impact. The interaction metrics (U.4.2 and U.4.3) obtain the maximum impact, being also quite high the reputation metrics (U.3.4 and U.3.6) and the posting activity in the stream (U.1.1). The most similar profile to influencers is that of employee (E), which mainly comprises community managers. However, this profile shows much lower quality in reputation metrics than influencers.

Finally, the TQC assesses the quality of the specific analytical tasks (topics) chosen by the experts. This cube captures the quality indicators that can be associated to the subset of facts represented by each topic. As an example, we have selected five analytical topics of the two domains at hand. We have included some fake or non-relevant topics that mainly correspond to memes or noisy expressions. Table 9 shows the statistics of the corresponding TQCs for the two domains. Notice that non-relevant topics always poorly cover the reference users in contrast to true topics. At this point, we should reject all topics with low coverage for reference users (shadow boxes in Table 9). The topic "Vendo Opel Corsa" is an example of noisy expression that have two different meanings: the literal sense "Opel Corsa for sale" and the ironic expression "it matters little to me". In this case, the latter usage prevails, resulting in low coverage of reference users. Notice also that the "natural disaster & tourism" data stream contains the false topic "Gran Turismo", which corresponds both to a car model and a videogame. This topic ranks second in terms of the number of tweets in this stream, contributing to a high level of noisy data. Considering both domains, we can see that the threshold to consider a topic as of low coverage for reference users is very different in each case. For the automotive domain it can be set at 1% (shadow boxes in the left part of Table 9), whereas for the second domain the threshold is in 0.1%, Notice that the coverage values for the reference users in the second domain are lower because the number of users of reference is smaller (7386 vs. 28,165).

Notice that TQC also gives us clues about the coverage of interesting topics of the data stream. Regarding the "natural disasters & tourism" domain, we can see that the topic "Cyclone Idai" has a low coverage in this data stream. This is due to how the data stream has been defined, which do not include any keyword related to specific disasters or events. To increase the coverage of these topics we need to redefine the set of keywords in a similar way than in the automotive domain.

Table 10 shows the results of the TQC for these topics taking as main quality criteria the coherence of the user profiles and posts (U2.1 and P2.1). That is, we rank Q-facts according to these two criteria and then evaluate the corresponding AP_{rel} scores. As expected, all topics deemed as relevant have as main relevant voices the professional and the journalist categories.

Table 9 Statistics of the topic examples used for the TQC (* *fake/non-relevant topics*)

Automotive Domain				Natural Disasters & Tourism Domain			
Topics	#Tweets	#Users	%Refer. Users	Topics	#Tweets	#Users	%Refer. Users
Car recalls	10,603	5,523	7.9%	Migrants	176,798	104,709	0.2%
Car repair	16,771	2,834	5.6%	*Gran Turismo**	106,282	59,152	0.03%
Sell accessories	108,537	4,820	2.3%	Tourism industry	85,966	55,777	1.2%
Stolen cars	2,381	1,787	2.3%	Disaster relief	83,835	59,978	0.8%
New models	722	486	22%	*Recipe for disaster**	76,102	68,818	0.08%
*Vendo Opel Corsa**	3,674	2,604	0.3%	Disaster management	44,403	29,256	1.3%
*Harrison Ford**	455	248	0.4%	Disaster response	25,558	17,471	2%
*ADHD vs. focus joke**	253	246	0.4%	Cyclone Idai	4,139	3,309	1.7%

Table 10 TQC for the relevant topic examples (%Relevant > 1) in the automotive domain

Analysis Topics (U2.1 and P2.1)	All categories	Influential users	User categories				
			Business users		Interested users		
			E	P	PS	J	L&F
Car recalls	0.613	0.633		0.598	0.190	0.516	
Car repair	0.866	0.224	0.166	0.875	0.264	0.156	
Sell accessories	0.673	0.877	0.103	0.768		0.398	0.542
Stolen cars	0.911	0.680		0.926		0.684	
New models	0.527	0.562	0.189	0.516		0.468	

5.3 Quality Indicators

Tables 11 and 12 present the results of the values of the quality indicators for users and posts respectively. Examining Table 11, we can deduce that a few users post most of the information (metric "on domain"), with a good degree of interactions and an acceptable coherence of their profiles. Regarding the reputation perspective, we can conclude that posting users are not influential but receive a good number of interactions (metric U4.2).

Table 12 reports the scores for usefulness and completeness of the automotive domain. It is noteworthy that only 31% of posts in the stream involve business dimensions. Most of these posts express associations between sentiment words and dimension values (29% of total posts). To measure completeness, we focus on a particular topic (e.g., car recalls) and measure the coverage of the different cuboids of interest for this topic. Table 12 shows that most of the posts in the topic at least mention the model or the brand. Facts involving all dimensions are covered by 30% of the posts in the topic.

Tables 13 and 14 report the quality indicators obtained for the "natural disasters" domain. In this case, the stream exhibits slightly lower user credibility compared to the previous domain but significantly higher user reputation. It is worth mentioning that this stream involves the main disaster relief organizations and tourist destinations. Regarding the Table 13, we can see much lower scores than in the automotive domain, indicating that most posts in this stream are not relevant to the analysis goals. For measuring completeness, we chose the topic "natural disaster management", which covers posts related to damage, location and main organizations involved in the recovery. Table 14 reveals that even when narrowing down the dataset to the specific topic of interest, the results are substantially worse than in the automotive scenario, indicating lower data quality. In this scenario, it makes no sense to measure the coverage of facts with polarity since many words involved in the dimensions have a negative polarity (e.g., disaster, victims, damage, etc.)

5.4 Filtering by Ranking

As a final step, we need to filter out low-quality data from the selected topics of analysis. In this case, we aim at identifying the main voices of the topic and then apply the best quality criteria to them. Let us illustrate this process with the analytical topic "car recalls" from the automotive domain. In this topic, we want to analyze car brands and models affected by recalls due to known manufacturing defects. The TQC in Table 10 shows us that the main business roles for this topic are Professional and Journalist. Tables 7 and 8 show us that the best criteria for these roles are P1.5 and U4.2 for journalist, and P2.2 and U2.1 for professional. Moreover, in the column Cut of Table 11, we can get the thresholds for some of these metrics.

Once the thresholds are applied, we can then rank the remaining facts according to the previous chosen criteria. In this case, we apply the criteria in sequential order from highest to lowest relevance to obtain the final classification. Finally, we set up a cut-off point to reject low-quality facts for this topic. As an example, Table 15 shows the 4-top and

Table 11 User quality indicators for the automotive domain

Perspective	Metric	Impact	Q Value	Coverage	Cut	Score
Credibility	U1.1. #Tweets in domain	0.749	0.871	0.776	> 6	**0.61**
	U2.1 Coherence	0.759	0.763	0.747	< 17	0,57
	U4.3 Total interactions	0.596	0.790	0.47	> 11	0.35
Reputation	U3.4 #Followers	–	–	–	–	–
	U4.2. #Received actions	0.595	0.897	0.41	> 9	**0.33**

Table 12 Post quality indicators for the automotive domain

Perspective	Dimensions	Coverage
Usefulness	All dimensions	31%
	All dims. & Polarity	29%
Completeness (Topic "car recalls")	Model or Brand	**97%**
	Model & Part	49%
	Model & Defect	44%
	Model & Part & Defect	30%

Table 13 User quality indicators for the "natural disasters" domain

Perspective	Metric	Impact	Q Value	Coverage	Cut	Score
Credibility	U1.1. #Tweets in domain	0.655	0.827	0.782	> 8	0.53
	U2.1 Coherence	0.733	0.791	0.628	< 26	0.51
	U4.3 Total interactions	0.692	0.814	0.777	$>$, 20	**0.55**
Reputation	U3.4 #Followers	0.807	0.63	0.61	> 320	0.50
	U4.2. #Received actions	0.822	0.78	0.541	> 1	**0.53**

Table 14 Post quality indicators for the "natural disasters" domain

Perspective	Dimensions	Coverage (%)
Usefulness	All relevant dimensions	23%
Completeness	Natural Disaster	21%
	Damage	25%
	Natural Disaster & Damage	5.4%
	Natural Disaster & Organization	3.6%
	Natural Disaster & Location	12%

4-bottom ranked posts. Top positions feature news reports on recalls of various brands and models by qualified users, while bottom positions include comments and opinions related to recalls but not reporting them.

6 Conclusions

The research presented here focuses on assessing data quality in social business intelligence (SoBI) applications. In this paper, we aimed at defining an integrated multidimensional view to capture the impact of quality metrics for the different types of facts extracted from social networks. This integrated model encompasses the main aspects pointed out in the literature, namely: credibility, reputation, usefulness, and completeness. The main components of the proposed model cover the main facts included in SoBI applications, namely: posts, users, and topics. As a main novelty, we claim that users must be the keystone in quality assessment. Therefore, we introduce the user role dimension to better understand the origin of the data, and to measure its impact on the quality metrics. This claim is in accordance with previous work that recommends adding a

Table 15 Top and bottom ranked posts for the topic "car recall"

Posts in Top Positions
ford recalls transit vans for air bags … {lnk} {lnk}
nissan recalls nearly 640,000 u.s. cars: nissan pathfinder, rogue, infiniti jx35, qx60: nissan north america has issued two separate {dots}
ford escape, transit connect recalled for dimwitted dash: -ford is recalling certain 2014 and 2015 escape suvs and transit connect va {dots}
ford escape, transit connect recalled for dimwitted dash: -ford is recalling certain 2014 and 2015 escape suvs and transit connect va {dots}
Posts in Bottom Positions
i5gornascimento pll @i5gornascimento mitsubishi faz recall do pajero full para trocar {qmark} airbag mortal {qmark} no brasi {dots} {lnk}
oligarcs like mitsubishi distributor wil not get away w thr excuses on thr montero.duterte wil make thm pay,recall wil b mandatory
stp revenue theft wt nigeria recalls yaris,hilux ova faulty airbags.shell faces risks 4rm $1.1bn nigrian oil scndl
{lnk} more recalls {punct} #sellcar #cardealer #buymycar #dealerbid #usedcars #cardeals #e4drive

user-domain dimension to credibility assessment to enhance the understanding of users' interest (Abu-Salih et al. 2019; Arenas-Márquez et al. 2021).

We have carried out experiments over two different long-term data streams for two separate domains. These experiments show the usefulness of the approach in revealing the main quality aspects that characterize each of these data streams. The primary distinctions between these streams are attributed to their data collection methodologies. Whereas in the automotive domain we used a large set of keywords for each car model, in the "natural disaster & tourism" we just used a few abstract keywords. As expected, quality indicators show better scores for the automotive domain, but not in all of the aspects. Users in the automotive domain exhibit lower scores in reputation metrics compared to those in the other data stream. Finally, the Q-cubes of the multidimensional model allowed us to better understand the quality features of our datasets and propose effective ways to select topics and filter out low-quality data.

The proposed methodology allows us to address any other analysis domain following the steps designed for it. Firstly, analysts must define a reference set of users and the keywords for retrieving the posts of interest. Since this step is exploratory, it assumes some knowledge of the application domain to identify an initial set of relevant users and a good choice of retrieval keywords. Once these two elements are defined, the proposed method automatically constructs all the analytical facts and calculates the quality indicators along with their impact. By selecting the top impact quality indicators, low-quality data can be filtered out. The entire process can be then refined by updating the retrieval query's keywords and/or the reference set of users. These updates can be suggested after the analysts have inspected the resulting dataset in the previous iteration, as well as by applying previous domain knowledge. Thus, new relevant users can be identified, and incomplete dimensions may require further keywords in the retrieval query.

The main practical implication of this study is that the proposed method allows analysts to measure the quality of the processed social media data from different perspectives and considering the profiles of the users that generate the contents. We demonstrate that data quality heavily depends on the domain and topics at hand, requiring the combination of different metrics according to their impact, different thresholds, and different filtering criteria. Our analysis of metrics in relation to user categories has provided us with valuable insights into the characteristics of the generated data, enabling us to formulate more effective strategies for filtering high-quality data.

This proposal has several limitations that need further research. The first is related to the language models used to measure coherence, usefulness, and completeness. In this paper, we used a simple approach by just taking the word distributions of the dimension values (e.g., car models, defects, natural disasters, etc.) However, this approach will depend heavily on the richness of the available metadata for analysis. Therefore, in complex domains with scarce linguistic resources, we will need new methods to achieve accurate results. In future research, we intend to explore advanced methods that leverage semantic annotations (Berlanga et al. 2015; Lanza-Cruz et al. 2018) and apply NLP sentence encoders (Reimers and Gurevych 2019) to enhance the precision and reliability of quality assessment.

Another limitation of our approach is the reliance on ad-hoc methods to construct reference collections of relevant users for the selected domains. These methods, which predominantly depend on predefined rules applied to screen names, profile descriptions, and expert-curated external resources, can be resource-intensive and do not provide the scalability and reliability required. To overcome this limitation, we need to develop more automated and robust techniques. Our initial approach involves a user classification into business roles, a process that can be further refined and automated through the implementation of advanced NLP text classifiers (Nebot et al. 2018; Lanza-Cruz et al. 2023). This will facilitate the identification of relevant users aligned with specific analytical goals.

Finally, our future research agenda also includes exploring innovative approaches to combine quality metrics effectively, with the aim of maximizing the AP (average precision) metric. Methodologies such as fastAP, developed for image retrieval (Cakir et al. 2019), can be adapted to our domain. However, it is essential to note that fastAP requires a pool of negative examples, a challenge we intend to address by using the provided set of reference users.

Acknowledgements This research has been partially funded by the Spanish Ministry of Science under grants PID2021-123152OB-C22 and PDC2021-121097-I00 both funded by the MCIN/AEI/ https://doi.org/10.13039/501100011033 and by the European Union and FEDER/ERDF (European Regional Development Funds). We also would like to thank valgrAI (Valencian Graduate School and Research Network of Artificial Intelligence) foundation for their support.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

References

Abu-Salih B, Wongthongtham P, Beheshti S, Beheshti B (2015) Towards a methodology for social business intelligence in the era of big social data incorporating trust and semantic analysis. In: 2nd International conference on advanced data and information engineering. Springer, Heidelberg

Abu-Salih B, Bremie B, Wongthongtham P, Duan K, Issa T, Chan KY, Alhabashneh M, Albtoush T, Alqahtani S, Alqahtani A, Alahmari M, Alshareef N, Albahlal A (2019) Social credibility incorporating semantic analysis and machine learning: a survey of the state-of-the-art and future research directions. In: Barolli L et al (eds) Web, artificial intelligence and network applications. Springer, Cham, pp. 87–100. https://doi.org/10.1007/978-3-030-15035-8_87

Abu-Salih B, Chan K. Y, Al-Kadi O, Al-Tawil M, Wongthongtham P, Issa T, Saadeh H, Al-Hassan M, Bremie B, Albahlal A (2020) Time-aware domain-based social influence prediction. Int J Big Data 7, Article 10. https://doi.org/10.1186/s40537-020-0283-3

Alrubaian M, Al-Qurishi M, Alamri A, Al-Rakhami M, Hassan M, Fortino G (2019) Credibility in online social networks: a survey. IEEE Access 7:2828–2855

Amigó E, Carrillo-de-Albornoz J, Chugur I, Corujo A, Gonzalo J, Meij E, de Rijke M, Spina D (2014) Overview of RepLab: author profiling and reputation dimensions for online reputation management. In: Kanoulas E et al (eds) Information access evaluation. Multilinguality, multimodality, and interaction. https://doi.org/10.1007/978-3-319-11382-1_24

Aramburu MJ, Berlanga R, Lanza I (2021) Quality management in social business intelligence projects. In: Proceedings of the 23rd International Conference on Enterprise Information Systems, pp 320–327. https://doi.org/10.5220/0010495703200327. https://www.scitepress.org/Papers/2021/104957/104957.pdf

Arenas-Márquez F, Martinez-Torres R, Toral S (2021) Convolutional neural encoding of online reviews for the identification of travel group type topics on TripAdvisor. Inf Proc Manag 58(5). https://doi.org/10.1016/j.ipm.2021.102645

Arolfo F, Cortés-Rodriguez K, Vaisman A (2022) Analyzing the quality of Twitter data streams. Inf Syst Front 24(1):349–369

Baeza-Yates R, Ribeiro-Neto B (1999) Modern information retrieval. Addison-Wesley

Bansal P, Bansal R, Varma V (2015) Towards deep semantic analysis of hashtags. ECIR. https://doi.org/10.1007/978-3-319-16354-3_50

Berardi G, Esuli A, Marcheggiani D, Sebastiani F (2011) ISTI@-TREC Microblog Track: Exploring the use of hashtag segmentation and text quality ranking. https://trec.nist.gov/pubs/trec21/papers/NEMIS_ISTI_CNR.microblog.final.pdf. Accessed 15 Jul 2022

Berkani N, Bellatreche L, Khouri S, Ordonez C (2019) Value-driven approach for designing extended data warehouses. DOLAP. http://ceur-ws.org/Vol-2324/Paper25-NBerkani.pdf. Accessed 15 Jul 2022

Berlanga R, García-Moya L, Nebot V, Aramburu MJ, Sanz I, Llidó DM (2015) SLOD-BI: An open data infrastructure for enabling social business intelligence. Int J Data Wareh Min 11(4):1–28. https://doi.org/10.4018/ijdwm.2015100101

Berlanga R, Lanza-Cruz I, Aramburu MJ (2019) Quality indicators for social business intelligence. In: 6th International Conference on Social Networks Analysis, Management and Security, Granada, pp 229–236. https://doi.org/10.1109/SNAMS.2019.8931862

Birjali M, Kasri M, Beni-Hssane B (2021) A comprehensive survey on sentiment analysis: approaches, challenges and trends. Knowl-based Syst 226

Cai L, Zhu Y (2015) The challenges of data quality and data quality assessment in the big data era. Data Sci J 14, Article 2

Cakir F, He K, Xia X, Kulis B, Sclaroff S (2019) Deep metric learning to rank In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 1861–1870, https://doi.org/10.1109/CVPR.2019.00196

Chauhan U, Shah A (2021) Topic modeling using latent dirichlet allocation: a survey. ACM Comput Surv 54(7)

Choi J, Yoon J, Chung J, Coh B-Y, Lee J-M (2020) Social media analytics and business intelligence research: A systematic review. Inf Proc Manag 57(6). https://doi.org/10.1016/j.ipm.2020.102279

Crawford M, Khoshgoftaar TM, Prusa JD, Richter AN, Al Najada H (2015) Survey of review spam detection using machine learning techniques. J Big Data 2(23). https://doi.org/10.1186/s40537-015-0029-9

Cresci S, Di Pietro R, Petrocchi M, Spognardi A, Tesconi M (2015) Fame for sale: Efficient detection of fake Twitter followers. Decis Support Syst 80:56–71

Czernek A (2018) Social measurement depends on data quantity and quality. Millward Brown Dynamic Logic. https://cupdf.com/document/social-measurement-depends-on-data-quantity-and-2014-07-17-social-measurement.html. Accessed 15 Nov 2022

Duan Y, Zhimin C, Furu W, Ming Z, Shum H (2012) Twitter topic summarization by ranking tweets using social influence and content quality. In: Proceedings of the 24th International Conference on Computational Linguistics, pp 763–780. https://www.aclweb.org/anthology/C12-1047

Francia M, Gallinucci E, Golfarelli M, Rizzi S (2016) Social business intelligence in action. In: Nurcan S et al (eds) Advanced information systems engineering. Lecture Notes in Computer Science, vol 9694. Springer, Cham

Gallinucci E, Golfarelli M, Rizzi S (2015) Advanced topic modeling for social business intelligence. Inf Syst 53:87–106

García-Moya L, Kudama S, Aramburu MJ, Berlanga R (2013) Storing and analysing voice of the market data in the corporate data warehouse. Inf Syst Front 15:331–349. https://doi.org/10.1007/s10796-012-9400-y

Gioti H, Ponis S, Panayiotou N (2018) Social business intelligence: review and research directions. J Intell Stud Bus 8:23–42. https://doi.org/10.37380/jisib.v8i2.320

Goonetilleke O, Sellis T, Zhang X, Sathe S (2014) Twitter analytics: a big data management perspective. ACM SIGKDD Explor Newsl 16(1):11–20

Gröger C (2021) There is no AI without data. Commun ACM 64(11):98–108. https://doi.org/10.1145/3448247

Gupta A, Kumaraguru P, Castillo C, Meier P (2014) TweetCred: real-time credibility assessment of content on Twitter. In: Proceedings of the 6th International Conference on Social Informatics, pp 228–243. https://doi.org/10.1007/978-3-319-13734-6_16

Hammou B, Lahcen A, Mouline S (2020) Towards a real-time processing framework based on improved distributed recurrent neural network variants with fastText for social big data analytics. Inf Proc Manag 57(1). https://doi.org/10.1016/j.ipm.2019.102122

Han X, Wang L, Liu G, Zhao D, Xu S (2017) Occupation profiling with user-generated geolocation data. In: 2nd International

Conference on Knowledge Engineering and Applications, pp 93–97. https://doi.org/10.1109/ICKEA.2017.8169908
Hernandez M, Hildrum K, Jain P, Wagle R, Alexe B, Krishnamurthy R, Stanoi IR, Venkatramani C (2013) Constructing consumer profiles from social media data. In: IEEE International Conference on Big Data, pp 710–716. https://doi.org/10.1109/BigData.2013.6691641
Holsapple C, Hsiao S, Pakath R (2018) Business social media analytics: characterization and conceptual framework. Decis Support Syst 110:32–45. https://doi.org/10.1016/j.dss.2018.03.004
Hu S, Kumar A, Al-Turjman F, Gupta S, Seth S, Shubham, (2020) Reviewer credibility and sentiment analysis based user profile modelling for online product recommendation. IEEE Access 8:26172–26189. https://doi.org/10.1109/ACCESS.2020.2971087
Immonen A, Pääkkönen P, Ovaska E (2015) Evaluating the Quality of Social Media Data in Big Data Architecture. IEEE Access 3:1–1. https://doi.org/10.1109/ACCESS.2015.2490723
Johannesson P, Perjons E (2014) An introduction to design science. Springer, ISBN: 978–3–319–10632–8
Jöhnk J, Weißert M, Wyrtki K (2021) Ready or not, AI comes – an interview study of organizational AI readiness factors. Bus Inf Syst Eng 63:5–20. https://doi.org/10.1007/s12599-020-00676-7
Kaufhold M-A, Christian M (2020) Rapid relevance classification of social media posts in disasters and emergencies: a system and evaluation featuring active, incremental and online learning. Inf Proc Manag 57(1). https://doi.org/10.1016/j.ipm.2019.102132
Keegan B, Rowley J (2017) Evaluation and decision-making in social media marketing. Manag Decis 55:15–31. https://doi.org/10.1108/MD-10-2015-0450
Kimball R, Ross M (2013) The data warehouse toolkit, 3rd edn. Wiley, p 48. ISBN 978–1–118–53080–1
Kolajo T, Daramola O, Adebiyi A, Seth A (2020) A framework for pre-processing of social media feeds based on integrated local knowledge base. Inf Proc Manag 57(6). https://doi.org/10.1016/j.ipm.2020.102348
Lanza-Cruz I, Berlanga R, Aramburu MJ (2023) Multidimensional author profiling for social business intelligence. Inf Syst Front. https://doi.org/10.1007/s10796-023-10370-0
Lanza-Cruz I, Berlanga R, Aramburu MJ (2018) Modeling analytical streams for social business intelligence. Inform 5:33. https://doi.org/10.3390/informatics5030033
Lauriola I, Lavelli A, Aiolli F (2022) An introduction to deep learning in natural language processing: models, techniques, and tools. Neurocomput 470:443–456
Lee I (2018) Social media analytics for enterprises: Typology, methods, and processes. Bus Horiz 61(2):199–210. https://doi.org/10.1016/j.bushor.2017.11.002
Lin J, Snow R, Morgan W (2011) Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 422–429. https://doi.org/10.1145/2020408.2020476
Nebot V, Rangel F, Berlanga R, Rosso P (2018) Identifying and classifying influencers in Twitter only with textual information. In: Nat Lang Proc Inf Syst 28–39. https://doi.org/10.1007/978-3-319-91947-8_3
Pääkkönen P, Jokitulppo J (2017) Quality management architecture for social media data. J Big Data 4(6). https://doi.org/10.1186/s40537-017-0066-7
Pasi G, Viviani M, Carton A (2019) A multi-criteria decision making approach based on the Choquet integral for assessing the credibility of user-generated content. Inf Sci 503:574–588. https://doi.org/10.1016/j.ins.2019.07.037
Păvăloaia V, Anastasiei I, Fotache D (2020) Social media and e-mail marketing campaigns: symmetry versus convergence. Symmetry 12(12):1940. https://doi.org/10.3390/sym12121940
Plachouras V, Stavrakas Y, Andreou A (2013) Assessing the coverage of data collection campaigns on Twitter: a case study. In: Demey Y, Panetto H (eds) On the move to meaningful internet systems. OTM 2013 Workshops. Lecture Notes in Computer Science vol 8186. https://doi.org/10.1007/978-3-642-41033-8_76
Podhoranyi M (2021) A comprehensive social media data processing and analytics architecture by using big data platforms: a case study of Twitter flood-risk messages. Earth Sci Inform 14. https://doi.org/10.1007/s12145-021-00601-w
Reimers N, Gurevych I (2019) Sentence-BERT: Sentence embeddings using Siamese BERT-networks, In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing
Rodríguez-Vidal J, Gonzalo J, Plaza L, Anaya-Sánchez H (2019) Automatic detection of influencers in social networks: authority versus domain signals. J Assoc Inf Sci Technol 70:675–684. https://doi.org/10.1002/asi.24156
Ruhi U (2014) Social media analytics as a BI practice: current landscape & future prospects. J Internet Soc Netw Virtual Commun. https://doi.org/10.5171/2014.920553
Sabherwal R, Becerra-Fernandez I (2013) Business intelligence: Practices, technologies, and management. Wiley
Sadiq S, Indulska M (2017) Open data: Quality over quantity. Int J Inf Manag 37:150–154. https://doi.org/10.1016/j.ijinfomgt.2017.01.003
Salvatore C, Biffignandi S, Bianchi A (2021) Social media and Twitter data quality for new social indicators. Soc Indic Res. https://doi.org/10.1007/s11205-020-02296-w
Saroj A, Pal S (2022) Use of social media in crisis management: a survey. Int J Disaster Reduct 48. https://doi.org/10.1016/j.ijdrr.2020.101584
Shankaranarayanan G, Blake R (2017) From content to context: the evolution and growth of data quality research. J Data Inf Qual 8:1–28. https://doi.org/10.1145/2996198
Sikdar S, Kang B, ODonovan J, Höllerer T, Adah S (2013) Understanding information credibility on Twitter. In: International Conference on Social Computing, Alexandria, pp 19–24. https://doi.org/10.1109/SocialCom.2013.9
Stieglitz S, Dang-Xuan L, Bruns A, Neuberger C (2014) Social media analytics. Bus Inf Syst Eng 6:89–96. https://doi.org/10.1007/s12599-014-0315-7
Stieglitz S, Mirbabaie M, Ross B, Neuberger C (2018) Social media analytics – Challenges in topic discovery, data collection, and data preparation. Int J Inf Manag 39:156–168
Tilly R, Posegga O, Fischbach K, Schoder D (2017) Towards a conceptualization of data and information quality in social information systems. Bus Inf Syst Eng 59:3–21. https://doi.org/10.1007/s12599-016-0459-8
Viviani M, Pasi G (2017) Credibility in social media: opinions, news, and health information – A survey. WIREs Data Mining Knowl Discov 7(5). https://doi.org/10.1002/widm.1209
Zachlod C, Samuel O, Ochsner A, Werthmüller S (2022) Analytics of social media data – State of characteristics and application. J Bus Res 144:1064–1076. https://doi.org/10.1016/j.jbusres.2022.02.016
Zhang R, Indulska M, Sadiq S (2019) Discovering data quality problems. Bus Inf Syst Eng 61:575–593. https://doi.org/10.1007/s12599-019-00608-0
Zheng L (2021) The classification of online consumer reviews: a systematic literature review and integrative framework. J Bus Res 135. https://doi.org/10.1016/j.jbusres.2021.06.038

Chapter 6

Social Media Multidimensional Analysis for Intelligent Health Surveillance

Publication

Aramburu, María José; Berlanga, Rafael and Lanza Cruz, Indira Lázara. Social Media Multidimensional Analysis for Intelligent Health Surveillance. Int. J. Environ. Res. Public Health 2020, 17, 2289. https://doi.org/10.3390/ijerph17072289. (Q2)

International Journal of
Environmental Research and Public Health

MDPI

Article

Social Media Multidimensional Analysis for Intelligent Health Surveillance

María José Aramburu [1,*], Rafael Berlanga [2] and Indira Lanza [2]

[1] Departamento de Ciencia e Ingeniería de los Computadores, Universitat Jaume I, E-12071 Castellón de la Plana, Spain
[2] Departamento de Lenguajes y Sistemas Informáticos, E-12071 Castellón de la Plana, Spain; berlanga@uji.es (R.B.); lanza@uji.es (I.L.)
* Correspondence: aramburu@uji.es

Received: 19 February 2020; Accepted: 26 March 2020; Published: 28 March 2020

Abstract: *Background*: Recent work in social network analysis has shown the usefulness of analysing and predicting outcomes from user-generated data in the context of Public Health Surveillance (PHS). Most of the proposals have focused on dealing with static datasets gathered from social networks, which are processed and mined off-line. However, little work has been done on providing a general framework to analyse the highly dynamic data of social networks from a multidimensional perspective. In this paper, we claim that such a framework is crucial for including social data in PHS systems. *Methods:* We propose a dynamic multidimensional approach to deal with social data streams. In this approach, dynamic dimensions are continuously updated by applying unsupervised text mining methods. More specifically, we analyse the semantics and temporal patterns in posts for identifying relevant events, topics and users. We also define quality metrics to detect relevant user profiles. In this way, the incoming data can be further filtered to cope with the goals of PHS systems. *Results:* We have evaluated our approach over a long-term stream of Twitter. We show how the proposed quality metrics allow us to filter out the users that are out-of-domain as well as those with low quality in their messages. We also explain how specific user profiles can be identified through their descriptions. Finally, we illustrate how the proposed multidimensional model can be used to identify main events and topics, as well as to analyse their audience and impact. *Conclusions:* The results show that the proposed dynamic multidimensional model is able to identify relevant events and topics and analyse them from different perspectives, which is especially useful for PHS systems.

Keywords: health surveillance; social network analysis; multidimensional analysis; text mining

1. Introduction

Public Health Surveillance (PHS) is defined as the ongoing systematic gathering, analysis, and interpretation of data, closely integrated with the dissemination of these data to the public health practitioners, clinicians, and policy makers responsible for preventing and controlling disease and injury [1]. In the last decade, many experiments have demonstrated that social media data can help public health officials to detect potential outbreaks, forecast disease trends, monitor emergency situations and gauge disease awareness and reactions to official health communications [2,3].

Traditionally PHS relied on surveys and primary data collected from healthcare providers and pharmacists on a weekly or monthly basis. While social media platforms cannot replace formal data sources for disease surveillance, they can provide complementary information with some advantages. Social media is a source of health and lifestyle information that covers all the society statements and is easy, rapid and cheap to obtain. However, for PHS, the filtering, integration and analysis of large amounts of data is as important as data gathering. The processing of social media data into

useful information for public health action is a complex task and a main research topic [4]. In general, current approaches address specific problems, and propose powerful ad-hoc solutions for analysing information stored in isolated and static repositories of social data [3]. Consequently, their algorithms and results cannot be applied to data from different contexts. However, PHS agencies require the development of general tools with advanced mechanisms for extracting, integrating and analysing social media streams with a wide range of different purposes [2].

The main objective of our research is to provide PHS agencies with a new tool to analyse the evolution of posts from several positions of interest. Its main advantage is to make it easier to learn who is posting, about what and with what purpose, or any combination of these. This tool considers social media sites as continuous sources of relevant posts that are processed in real time to produce useful data. Then, users can execute operations to analyse the incoming data in order to monitor trends, and quickly identify events and alarms. Online analysis tasks consist of summarising the data at different levels of detail and from complementary points of view, depending on the purpose and the scope of the research. Furthermore, when needed, the pre-processed input data can be integrated with other data sources or stored in sandboxes dedicated to statistical analysis and historical studies. With this tool, PHS services can turn large amounts of posts into meaningful information, and easily and understand what is happening in social media in a timely manner.

2. Previous Work

By analysing the latest advances of the Centers for Disease Control and Prevention, the report presented in [1] helps to understand the current state of development of PHS systems. In general, current systems present enhancements for data gathering (mainly, the integration of different data sources), data analysis and visualisation (mainly, online available data repositories), and dissemination (mainly, specialized, periodically published reports). However, these systems rely on disease-specific approaches that inhibit efficiency and interoperability and use out-of-date technologies that no longer meet user needs for data management, analysis, visualization, and dissemination. A main recommendation of [1] is that advances in information technology, data science, analytic methods, and information sharing provide an opportunity to enhance surveillance. Social networks as a new source of data for PHS were not mentioned in this review, showing that current systems are in their early research stage.

2.1. Using Social Networks for PHS

In the literature, many works demonstrate the usefulness of social media data for PHS. A research review presented in [5] already confirmed that social networks can be used as platforms to track the spread of infectious diseases, as well as warning detection systems with faster response than official data, which can take one or two weeks to collect and process. The authors conclude that PHS based on social media will never replace traditional surveillance, but social networks provide complementary data to be integrated with traditional sources. In their review, apart from the large amount of noisy data, the main open issues are the lack of coverage and the bias in the collected social media data. As solution, they propose new filtering methods for data quality together with advanced frameworks to integrate data from different social media platforms.

In a complete and recent review [3], Twitter data has been found to be useful for several different public health applications, including monitoring diseases, public reactions, outbreak/emergency situations, prediction, lifestyle, geolocation, and other general applications. Authors highlight that most work done to date is not applicable to many use cases. They claim for a system that can be applied across researchers to categorize tweets in real-time is very important in order to better track information and facilitate rapid decision making. As in [5], the lack of coverage and the bias in the collected social data are the main drawbacks. However, in contrast to [5], they also consider that users are more important than tweets in public health research and recommend future studies to switch their unit of analysis from tweets to individual users.

As reviewed in [3], many previous approaches successfully apply Machine Learning (ML) tools in order to manage large amounts of data at tweet level. Most of them use supervised ML methods that need manually annotated data collections. These systems classify posts in several ways, as for example: (i) cleaning noisy data with adverse drug reactions (ADR) [6] or with medical conditions [7]; (ii) observing the frequency and distribution of textual ADR mentions [8], and identifying ADR semantically annotated data [9]; (iii) grouping tweets by topic [10]; and (iv) assigning a polarity to both tobacco-related tweets [11] and posts about personal sentiments [12] or influenza [13].

In the literature, there are also some papers that apply Text Mining for identifying tweets' contents, such as topic modelling and Latent Dirichlet Allocation (LDA) [3]. For example, in [14], a new associative topic model that identifies relevant tweets by using a combination of keywords and associated topics, obtained good results for monitoring diseases. In [15], unsupervised LDA-based clustering using topic modelling performed well for detecting relevant Twitter data. Compared to a trained classifier, the clustering method was found to offer less control over the topics, but the classifier was costlier because it required many manual annotations. Another LDA method of analysis presented in [16] was applied to large amounts of tweets in order to find topics related to public health. Their research results showed that only very common topics were detected.

Following the recommendations in [5], it is important to note that many efforts have focused on processing the posts independently and ignore the fact that the value of a post is inherently related to the credibility of its poster. At this level of analysis, the paper in [17] proposes a ML framework to classify social media users, which relies on data from user profile accounts, user tweeting behaviour, textual contents, and social network parameters. The framework was successful in identifying political affiliations, but failed in identifying user races. Similarly, in [18], an approach for enhancing PHS with quality social media data coming from trustworthy users is presented. This work proposes six trust filters to rank social media users with respect to a given criteria. Preliminary results show that the best filters are based on the number of related posts a user sends. The profiling of Twitter's users to enhance tweet classification and relevance was already proposed as an open issue in [19].

2.2. Multidimensional Analysis of Tweets

In this section, we review some few works with similar purposes to ours, that is, works that process tweets in order to support multidimensional analysis.

In [20], authors propose an active surveillance methodology based on four dimensions: volume, location, time and public perception. They explore the public perception dimension by performing sentiment analysis and a clustering approach is used to exploit the spatio-temporal dimensions. Authors show that Twitter could be used to predict, spatially and temporally, dengue epidemics. They also propose a dengue surveillance approach that produces a weekly overview of what is happening in each city compared to the weeks before.

The M-Eco system, presented in [21], collects data from social media and TV/radio for public health monitoring purposes. It allows a user to search for disease names or symptoms and to assess the related signal information by means of a geographic map, a tag cloud or a timeline. Previously, texts are automatically annotated to identify diseases, person and location names. ML algorithms detect patterns in the data and, since the large number of generated signals can overwhelm a user, recommendation techniques are exploited to filter out those signals that are of potential interest for a user. Finally, the information is shown in charts and through personalized tag clouds to allow users to easily assess signals. Together with the frequent changes in posts terminology, again, the lack of social coverage and noisy data are identified as main limitations.

Considering the spatial dimension of tweets, in a recent work [22], Twitter data are used to extract spatio-temporal behavioural patterns to monitor flu outbreaks and their locations. In addition, they propose future extensions to study the epidemic spread of flu within different subpopulations by leveraging socio-economic and demographic data. Their results verify that flu-related traffic on social media is closely related to actual flu outbreaks. More specifically, they also find that clinical

flu encounters lag behind online posts, and identify several public locations from which a majority of posts initiated. The main limitations of this study are very similar to previous works. The large amount of noisy data requires advanced classification techniques or machine learning approaches for deeper content analysis of social media posts. Furthermore, because not all users have their GPS enabled or declare their location in their social media profiles, spatial analysis of social media may be biased. The lack of GPS labels in posts is an open problem that some systems have already tried to solve by processing posts contents and metadata [7,10,23,24].

2.3. Modern Related Technologies

In our work, with the purpose of analysing social media contents from several positions of interest, we need to process the relevant posts in order to produce useful multidimensional attributes. Then, these parameters can be applied to summarise tweets and posters' data at different levels of detail and to analyse social networks data from complementary points of view. As we have seen, none of the previous works have proposed the development of a system with this functionality. Such a solution requires the application of advanced information technologies. This section reviews the main technologies that our approach takes advantage of. More specifically, it reviews the following technologies: Business Intelligence, data quality, intelligent processing of data, and finally, streaming technologies.

2.3.1. Business Intelligence

The main objective of Business Intelligence (BI) is extracting strategic knowledge from the information provided by different data sources to help companies during decision-making. The processing and analysis of massive data oriented to BI has evolved in recent years. Traditionally, the most commonly used approaches have combined data warehouse, online analytical processing (OLAP), and multidimensional design technologies [25]. OLAP tools were introduced to ease information analysis and navigation from large amounts of transactional data. These systems rely on multidimensional data models, which apply the fact/dimension dichotomy. Multidimensional data are represented as facts, whereas dimensions define hierarchies with different detail levels to aggregate data.

BI systems work on very specific scenarios, making use of static and well-structured data sources of corporate nature, and causing all the information to be fully materialized and periodically processed in batch mode for future analysis. More recently, new technologies for Exploratory OLAP have been introduced, aimed at exploiting semi-structured external data sources (e.g., XML, RDF) for the discovery and acquisition of relevant data that can be combined with corporate data [26].

Social networks have become a new source of valuable information for companies, helping them, among others, to know the opinions of their customers, to analyse market trends, and to discover new business opportunities [27]. The main purpose of Social Business Intelligence (SBI) is to help managers in making decisions by performing a multidimensional analysis of the relevant information disseminated on social networks. SBI and OLAP tools can enable the definition of hierarchies of analysis that organise and classify posts and users from complementary points of view and at different levels of detail [28,29]. Thus, from our point of view, by processing the posts relevant for PHS tasks, it is possible to build the multidimensional metadata structures applied by OLAP operators to aggregate social media data. At the end, BI systems can allow analysts to obtain many different visual representations that summarise the reality expressed in a large amount of posts.

2.3.2. Data Quality

When dealing with social media data, the quality of data becomes a main issue [30]. Users and posts require measures to assess their usefulness for health surveillance. Measures such as the number of replies, likes, or retweets are indicators of the relevance of the posts. Combining them with parameters expressing the credibility of the sending user (e.g., number of followers or verified account) is also useful for cleaning up social media data. Posts without meaning or acceptance, as well

as posts coming from users out of context or without credibility should be considered noisy data and be discarded by the system.

Furthermore, as pointed out by [3], retrieving relevant posts using a list of keywords might be problematic. The reason is that many tweets are relevant, but do not mention the predefined words, whereas many tweets including the keywords may be irrelevant for different reasons (misunderstandings, bots, spams, etc.). The main conclusion is that, after defining an initial list of keywords, the quality of the retrieved collection needs to be analysed with two main purposes, first, to find the best group of keywords, and second, to remove noisy posts [28,31,32].

In the literature, quality measures are defined at post and user level. At post level, there are many metrics covering the characteristics of the text (e.g., grammar, contents and semantics), together with the metrics specific to micro-blogs that reflect their social impact (e.g., number of retweets). On the other hand, at user level, there are activity metrics to assess the posters' relevance (e.g., account age and number of posts) and popularity (e.g., number of followers, likes and mentions).

The evaluation of the quality of contents published on micro-blogging platforms has focused mainly on post-retrieval operations. Searching for posts related to a topic [33,34]; filtering posts based on their credibility and quality [35,36]; detection of events and disasters [37–39]; analysis of feelings, political and consumer opinions [40–42]; and the detection of influencers [43,44], are some example applications. Other applications aimed at the detection of spammers, bots and advertising campaigns have proposed intelligent analysis techniques for social metrics [45–47].

2.3.3. Intelligent Processing of Data

Entity resolution, topic classification and user profiling are intelligent processing tasks to extract meaningful data and semantically annotate posts. There are many types of meaningful data, such as the subjects and topics of posts, user affection or expertise, as well as post intent, and text polarity [29]. These automatic processes may apply different complex techniques [48]: crawler meta-data, information extraction, information retrieval, natural language processing, machine learning, and domain knowledge resources [4].

In recent years, we have witnessed a great interest in massively annotating the biomedical scientific literature [28]. Most of the current annotators rely on well-known lexical/ontological resources such as MeSH, Uniprot, and UMLS. These knowledge resources usually provide both the lexical variants for each inventoried concept and the concept taxonomies. Most semantic annotation systems are dictionary look-up approaches, that is, they rely on the lexicon provided by the ontology in order to map text spans to concept lexical variants. Although in a simpler way, the same techniques can be applied to find the topics associated to PHS-relevant posts.

In social media applications, opinion mining and sentiment analysis have been important research areas that combine techniques from Machine Learning (ML) and Natural Language Processing (NLP). One of the most relevant applications of sentiment analysis is aspect-based summarization [49]. Given a stream of opinion posts, aspect-based summarization is aimed at extracting the most relevant opinionated aspects along their sentiment orientation, usually represented as a score and a polarity. Aspect-based summarization has been usually divided into three main tasks, namely: sentiment classification, subjectivity classification and aspect identification. The first one is focused on detecting the sentiment orientation of a sentence, the second one consists of detecting if a sentence is subjective (i.e., if it contains a sentiment), and the latter one consists of detecting the most relevant aspects of an opinion stream. ML-supervised approaches have been widely adopted to solve these problems, because they can be easily modelled as traditional classification problems. Unfortunately, it is unfeasible to get training examples for all the items and potential aspects regarded in opinion streams. Thus, supervised approaches have been restricted to obtain sentiment lexicons and to detect sentence subjectivity with them [49]. Sentiment analysis in open scenarios should rely on unsupervised or semi-supervised methods [50]. Moreover, for social media data, sentiment analysis must be combined with social

network parameters, which measure the diffusion and popularity of opinions spread across social networks [51].

2.3.4. Streaming Technologies

The methods and architectures for BI are evolving. Traditional architectures consist of large data warehouses that integrate various data sources into a data repository under a multidimensional scheme. Data processing is executed in batch, which causes late alerts and delays in decision-making. A newer approach, in accordance with the current needs for Big Data processing, focuses more on the speed and immediacy of information, processing data in streaming and, when needed, building sandboxes where to execute batch analysis processes. In this way, only the data items needed for the knowledge models are stored, optimizing memory usage.

Modern software architectures and programming technologies for real-time processing can be applied to process posts as soon as they occur. Some authors [29,52] propose an extended Lambda Architecture for Big Data processing [53] that includes semantic data processing and establishes mechanisms to semantically enrich raw data with metadata from various sources. Furthermore, it is necessary to adapt the hierarchies of analysis and the semantic processing of post to the dynamic behaviour of social media sites. This requires mechanisms to discover and add new aspects of analysis on the fly.

2.4. Conclusions

Summarizing this review, there are some unresolved problems that limit the utilization of social media data for PHS, namely: social coverage, bias in the available data, poor quality and noisy data, lack of information about users, language and multilingualism, etc. However, by applying computational intelligence and modern technologies, a large amount of useful information (e.g., topics, news, needs, events, sentiments) can be extracted from social media, to be processed and delivered for many applications. In [2], three major applications for social media in PHS are identified: epidemiologic monitoring, situational awareness during emergency response and communication surveillance. Therefore, as a general conclusion, we consider that it is time to provide health officers with intelligent tools prepared for the multidimensional analysis of social media data about a wide range of ailments, and with interactive functionalities for many surveillance tasks.

3. Methods

Although not everyone shares health conditions in social networks, they are so widespread that their study can serve PHS officers as a complement to traditional data sources [21,54]. The relevant aspects of social media posts for health are the following ones:

- *Sender users*. Active users in social media can be individuals (i.e., personal accounts that range from health professionals to lay users) as well as organizations (i.e., accounts belonging to associations, laboratories, official bodies, etc.). Lay users post mainly about their symptoms, diagnosis, announcements, opinions, and sentiments. However, collective users representing their organizations post journal news, recommendations, official announcements, events, promotions, etc. Organizations have a mission and a specialty. For example, the user named Cancer Discovery sends tweets with high-impact articles and news on major advances in cancer research, and the World Health Organization is committed to promoting good health habits. For organizations, the activity level in social media (i.e., number of posts and social media campaigns) and the credibility (i.e., number of followers and verified accounts) is usually much higher than for individuals. From a PHS perspective, individuals could be classified by degree of affectation (as patients, familiars, or interested) as well as by degree of knowledge (as health masters, specialists or professionals);
- *Posts contents*. The words in the posts can be applied to retrieve and filter them, but also to annotate them with the topics that describe their contents at different levels of detail. Topics can

range from the most general subjects (i.e., medical condition, health campaign, financial service, etc.), to general topics (i.e., skin disease, WHO alert, AIDS treatment, etc.) up to very specific topics (i.e., nodular melanoma, sun protection recommendations, etc.). Real-world named entities, such as persons, locations, organizations, ailments and drugs, are also annotation elements that can be automatically identified. Post metadata elements like user locations, hashtags (# symbol) and screen names (@ symbol) should also be considered when annotating posts' contents;

- *Posts polarity*. Many people choose social media to express positive and negative sentiments and opinions. Processing post contents into polarity values is useful to analyse social awareness and acceptance, or to detect alarms and conflicts in the population [19];
- *Posts temporality*. Users post their news as soon as they occur, so that social media is a source of timely information for PHS. The posts refer mainly to health issues that may occur uniformly and continuously, as well as seasonally or due to outbreaks. The temporal dimension of posts could allow researchers to filter them as well as to group them into different time units (i.e., daily, weekly, monthly, etc.). This would make easier to monitor the evolution of real facts, to detect reactions to announcements, to discover new information, and even to filter noisy data. For example, a peak in the distribution curve of posts about melanoma can serve to discover a huge amount of unusable posts reacting to the announcement of a famous person suffering from it.

In our approach, the intelligent processing of incoming posts automatically generates semantic annotations to represent different aspects. At the same time, this processing also allows for the filtering and classification of large amounts of social media data. Then, by combining one or more of these aspects, analysts can group the posts and summarise them in different ways.

For example, PHS officers may be interested in monitoring public reaction to seasonal campaigns reporting on how to prevent melanoma. After retrieving the relevant posts by means of keywords, the resulting collection could be large and difficult to manage. With our approach, the advanced processing of these posts annotates them as sent by organizations, general lay users or patients, containing opinions with positive or negative polarity, etc. Then, by adding conditions on these aspects and on the temporality of the posts, it would be possible to obtain meaningful indicators that measure social reactions from different points of view.

Other aspects of social media users related to epidemiological and demographic studies include personal attributes such as race, age, sex, or geographic location. Lay users do not specify these data in their profiles; consequently, there have been many efforts to extract them from posts'contents [10,17,55]. Currently we do not regard these aspects in our research tools.

Multidimensional Analysis of Twitter Streams

Our approach relies on multidimensional data models dynamically defined over Twitter streams [31]. More specifically, for PHS, we propose the star-based schema shown in Figure 1. The main purpose of this schema is to generate summarized topics and events from the stream so that they can be analysed as required by PHS tasks.

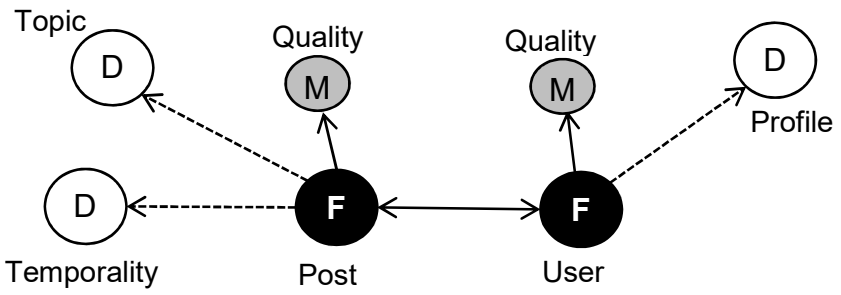


Figure 1. Summary of the multidimensional model for Twitter streams analysis. We follow the usual notation for multidimensional models: (D) for dimensions, (F) for facts and (M) for metrics. Dotted lines indicate that dimensions are dynamically created from the streamed data.

As can be observed, social media data elements are modelled with two different fact types. Post Facts consist of metrics extracted from post contents, manly topics, temporal attributes and other tweet quality parameters like number of replies or retweets. From a different perspective, User Facts include profile metrics extracted from descriptions and other user level quality parameters, like if the sending account is officially verified. Notice that some user metrics can be calculated by adding up the post metrics of all the tweets sent by the same user. The double-headed arrow connecting Post and User facts represents the binary correspondence that exits between the posts of a user and the poster of a tweet. This correspondence is needed to calculate some metrics, like the total number of replies of the relevant posts sent by a user.

In the proposed model, dimensions are dynamic in the sense that their values are derived from the incoming data. For example, the Topic Dimension accounts for the main subjects the tweets are discussing, which are statistically inferred after a certain number of tweets have been processed. More specifically, in this paper we propose a bigram-based temporal analysis to detect these topics in the stream of tweets. Furthermore, the temporal behaviour of the published information is quite relevant to characterizing posts. For example, a tweet can take part of a relevant news event, a health awareness campaign, or it can just be an isolated opinion. Thus, identifying the Temporality Dimension of posts allows us to properly group and analyse the information of the stream.

A relevant aspect included in our model is the quality of posts and users. Quality Dimension parameters can be derived from a combination of both, metrics provided by Twitter in the posts metadata (e.g., number of replies of tweet or the number of followers of a poster) and post contents metrics (obtained by processing the textual part of the posts). The resulting set of quality metrics indicates how reliable the information is with respect to the analysis at hand. For example, many users that publish a high quantity of posts can be spammers whose messages produce a great bias over some topics. Some messages can be just noise due to keyword ambiguity or more complex aspects like irony, sarcasm and metaphors. This last aspect is well known in the literature because some medical conditions like cancer and metastasis are often ironically used in other domains like politics. Quality metrics aim to assess how well the processed information fits with the analysis domain.

Finally, we include a Profile Dimension for users in order to identify their main intentions within the social network. Profiling users is very useful for social network analysis since it allows us to characterize both those who are posting information and those who are interacting with information. For PHS, we would like to dynamically distinguish between journalists, health-care professionals and services, as well as concerned people. Concerned people are individuals that have some true relationship with the health conditions being tracked (e.g., cancer survivors, relatives of patients). Profiling is performed over the brief text provided by the users in their Twitter accounts to introduce themselves. More specifically, we first identify and manually label a set of users by extracting meaningful frequent bigrams from their descriptions. Then, we train a neural-network classifier with the labelled users to predict the profiles of the incoming users.

In order to demonstrate the usefulness of the proposed streamed multidimensional analysis model, we designed a Twitter multi-keyword stream involving several topics from the PHS domain (see Table 1). The aim of the query is to reproduce a large and heterogeneous stream of tweets to monitor.

Table 1. Keywords for the Twitter stream for Public Health Surveillance.

Conditions: cancer, tumor, carcinoma, leukemia, lymphoma, metastasis, sarcoma, skin cancer, melanoma
Treatments: chemotherapy, homeopathy
Observations: UV, Cholesterol, LDL, MRI

For analysing users, we define two quality metrics according to the information they publish, namely: *domain coherence* and *vocabulary diversity*. Both metrics rely on the language model derived from the tweets, that is, the distribution of their words. For domain coherence, we calculate the joint distribution between the user's language model and a distribution of medical terms from UMLS Meta

Thesaurus®. In our experiments, UMLS achieved a great coverage of medical terms, including the most common forms used by lay users to refer conditions and diseases. For vocabulary diversity, we propose the Yule's I metric [56] applied to all terms (not only UMLS).

4. Results and Discussion

By using the Twitter API search method with the keywords shown in Table 1, we collected a large stream of tweets from June to December 2019. In this stream, a total number of 777,778 posts were tweets published by 395,804 users, where 77% of them posted just once in the stream. Concerning the overall quality of the stream, it is worth mentioning that the collected data are quite redundant as only 281,260 tweets consisted of unique texts, the rest being copies of other tweets. This fact shows how easily some streams could be biased towards the most repeated tweets. Finally, it is important to note that the retweets in the stream were treated separately, with the sole purpose of analysing user interaction.

4.1. Analysing Users

Figure 2 shows the user quality metrics for the top users according to the number of published tweets in the stream. Bots and spammers usually have a high number of tweets with a very small Yule's I score (lower than 30). Users with very low domain coherence and a high Yule's I score are usually prolific authors whose screen names contain some of the keywords of the stream, but their tweets are out of domain. By using these metrics, we can easily filter out non-relevant users (red dots in Figure 2). Notice that removing them implies rejecting around 15% of tweets from the original stream. Further details about the top users' metrics are shown in Table 2.

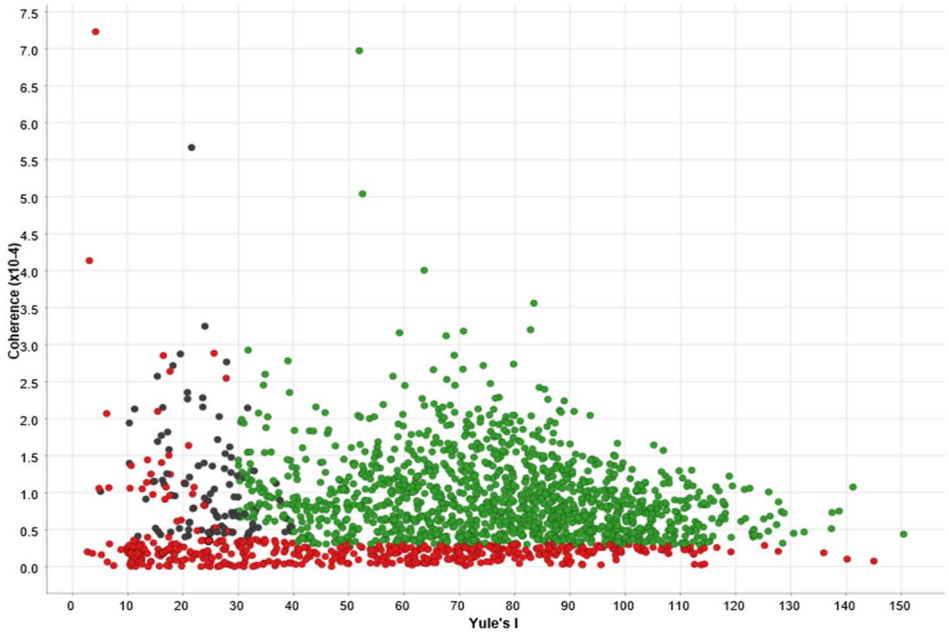


Figure 2. Quality metrics for the top-2000 users according to published tweets in the stream. Green points are good users, black points are users that need manual inspection, and red points are users that should be filtered out.

Table 2. Top-10 users' quality metrics. In the screen names column, stream keywords are in bold face. Shadowed rows correspond to out-of-domain users.

Screen Name (User)	#Tweets	Yule's I	Coherence (10^{-4})	Description
QunolOfficial	3306	42,3	0,88	Forum
ldl_bot	2556	94,3	0,21	Comics bot
LymphomaPapers	2110	75,4	1,39	Academic bot
treda10	1740	70,6	1,15	Company CEO
AvosFromMexico	1623	49,9	0,33	Food sales
EurekaMag	1194	71,2	1,87	Magazine
medvizor	1053	56,6	0,65	Forum
beechnutwx	1010	11,2	0,31	Weather bot
SnoMiddleForkWX	963	29,8	0,32	Weather bot
Sara_**sarcoma**	947	113,7	0,02	Influencer

4.2. Events and Topics

In the tweet stream, health-related events and topics are identified as bigrams grouped according to their temporal distributions of occurrence at day level. More specifically, we establish three thresholds, namely: a minimum kurtosis (10), a maximum peak of occurrences (20), and a maximum time span where the bigram occurs (10). As a result, around three thousand bigrams were selected. As these bigrams usually co-occur in many related tweets, we cluster bigrams by applying a graph modularity algorithm [57], resulting in 260 groups representing the full range of topics and events reported by the incoming tweets. It must be noticed that this process is fully unsupervised, and it aims at capturing the relevant events and topics of the stream. The detected events were manually inspected through web searches in order to confirm the results of the clustering method, and to demonstrate the usefulness of the quality metrics.

The impact of event-related tweets is around 7% of the total of tweets, whereas topics cover 51% of the incoming tweets. Additionally, 60% of event-related tweets have been also associated to some topic. Consequently, when analysing some topics, we must be aware of event-related tweets as they can have a high impact.

Tables 3 and 4 show the top frequent events detected for posts written in English and Spanish, respectively. Events are usually related to some media news involving famous people from politics or sports. In Spanish, many events are related to children whose cases become viral. In these tables, we use the following quality metrics: *Y* (average Yule's I), *%Ver.* (percentage of verified accounts) and *OnDom.* (average number of tweets per user in the stream). Furthermore, the last column shows the number of users tweeting and retweeting about these events.

Table 3. Top frequent events (English).

Event (bigram)	#Tweets	Main Topic	User's Quality (Y, %Ver., OnDom.)	#Users Tweet./Retw.
Alex Trebek	2659	Pancreas cancer	(70, 31%, 6)	2200/4535
Carlos Carrasco	659	Leukemia	(72, 28%, 4)	589/2473
Tapeworm woman	554	False tumor	(65, 39%, 6)	437/332
Ross Perot	357	Leukemia	(85, 23%, 4)	335/1093
Bank holidays (chemo canceled)	86	Chemotherapy	(68, 0%, 5)	84/27

Table 4. Top frequent events (Spanish).

Event (bigram)	#Tweets	Main Topic	User's Quality (Y, %Ver., OnDom.)	#Users Tweet./Retw.
Tabare Vázquez	921	Lung Tumor	(57, 27%, 6)	691/1521
Carlos Carrasco	854	Leukemia	(55, 33%, 5)	262/622
Francia homeopatía	338	Homeopathy	(59, 11%, 6)	293/638
Luis Enrique	282	Cancer	(71, 13%, 3)	102/1331
fármaco evita	108	Cancer	(61, 7%, 7)	103/84

The number of selected bigrams for topics is similar to that for events, around four thousand bigrams. These bigrams can be organized into 139 groups. Tables 5 and 6 show the main topics according to the most frequent bigrams associated to different types of cancer for English and Spanish, respectively. It is important to notice that some few topics like "leukemia" and "melanoma" have been defined with unigrams instead of bigrams as their semantics are unambiguously associated to one single word. It must also be said that topics contain those events with the corresponding topics, so we must be aware of the impact of these events in the topics (e.g., "leukemia" topic). Notice that the users involved in topics are usually more active than those involved in events, mainly due to awareness campaigns associated with these topics.

Table 5. Top frequent topics (English subset).

Topic	#Tweets	User's Quality (Y, %Ver., OnDom.)	#Users Tweet./Retw.	%Tweets (Replies, 1st_Person)
leukemia	27.000	(76, 6%, 12)	3.200/17.300	(16%, 12%)
melanoma	26.791	(78, 8%, 15)	2.700/3.364	(14%, 7%)
skin cancer	19.000	(72, 7%, 5)	8.219/23.022	(23%, 12%)
brain tumor	18.900	(70, 5%, 5)	6.091/11.268	(37%, 28%)
lung cancer	1.500	(75, 4%, 45)	401/396	(18%, 8%)

Table 6. Top frequent topics (Spanish subset).

Topic	#Tweets	User's Quality (Y, %Ver., OnDom.)	#Users Tweet./Retw.	%Tweets (Replies, 1st_Person)
leucemia	12.390	(57, 8%, 4)	1.388/11.108	(23%, 7%)
tumor cerebral	4.451	(67, 5%, 3)	1.264/9.951	(23%, 16%)
melanoma	3.726	(60, 15%, 9)	208/421	(8%, 3%)
cáncer mama	2.800	(58, 4%, 6)	221/188	(17%, 12%)
cáncer páncreas	1.100	(57, 3%, 4)	259/242	(2%, 3%)

Some interesting features we analyse are the use of first person and the percentage of replies associated with each topic (last columns of Tables 5 and 6). It is worth mentioning that around 4% of tweets in the topic melanoma are indeed related to politics. This fact has been easily detected by exploring the most frequent subtopics (bigrams) associated with melanoma. In this case, we find out that 70% of the corresponding tweets are indeed replies, in contrast to the percent of replies of the whole melanoma topic (14%). On the other hand, the use of the first person indicates the ratio of people directly affected by the diseases.

Finally, Figure 3 serves to analyse the authors (in blue colour) and audience (in red colour) of some topics and events grouped by user profiles. As explained in Section 3, authors are identified as the posters of the relevant tweets, whereas the audience includes those users interested in retweeting the same tweets. Then, according to the language models of their descriptions, users can be classified into different profiles. Table 7 shows the language models obtained for these profiles.

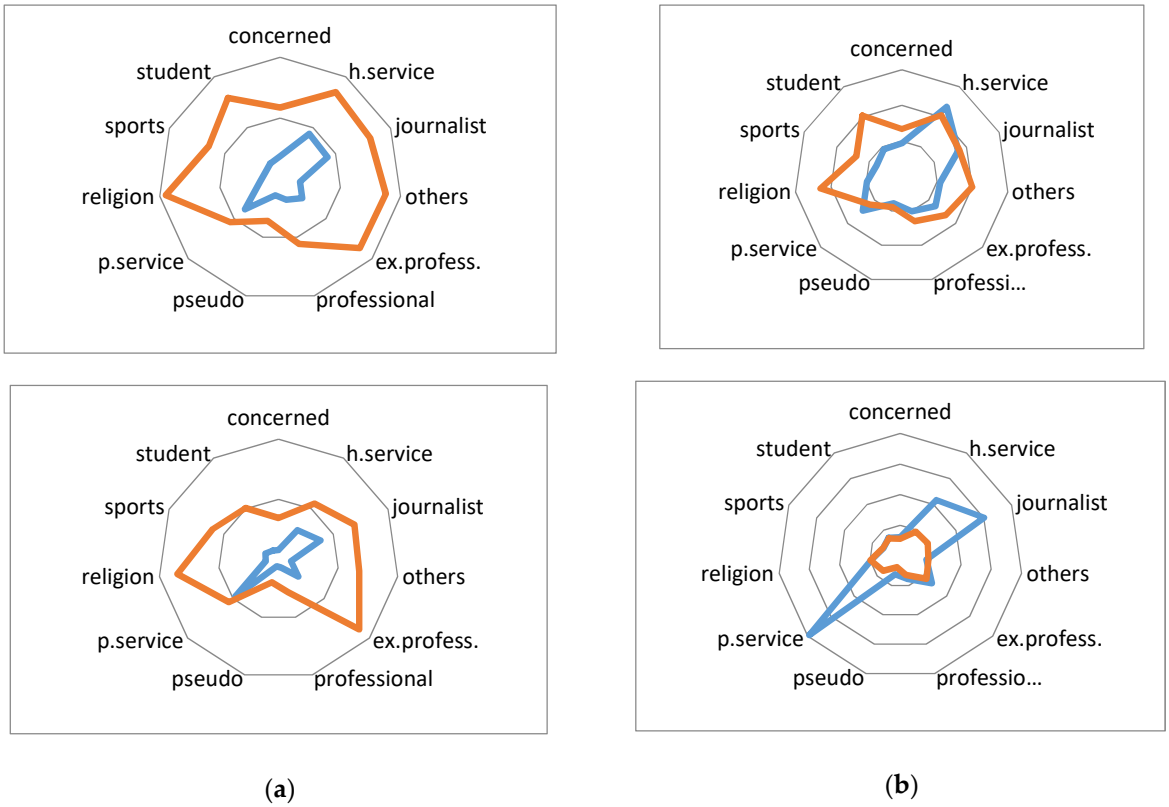


Figure 3. Comparing profiles from different topics and events (authors in blue and audience in red): (**a**) Top: Leukemia (topic), Bottom: Carlos Carrasco (event) (**b**) Top: Skin Cancer (topic), Bottom: Tapeworm (event).

Table 7. Language models for different profiles according to the users' descriptions.

Profile	Definition	Top Words in Users' Descriptions
professional	Specialists	radiosurgery, oncoplastic, oncologist, haemato-oncology, consultants
ex-professional	Retired people	retired, former, senate, viet, colonel
student	Students	student, thesis, undergraduate, engineering, studying
sports	Sportman	runner, marathon, athlete, skater, rider
religion	Religious terms	amin, allah, savior, hindu, jesus
p.services	Public services	traffic, 24h, breaking, weather, protection
h.services	Health-care services	urgencias, screenings, specialties, uninsured, visitors
concerned	Disease-concerned people	survivor, reminder_ribbon, survivorship, warrior
pseudo	Alternative therapies	reiki, meditation, yoga, ayurveda, remedial
journalist	Press and publications	medline-indexed, issn, journal, indexed, open-access
others	Other people	zombies, yin-yang, weekends, voracious, virgo (all with low probs)

Notice the different nature and impact of the profiles associated to each kind of event and topic. At the left side of Figure 3, the posts about the "leukemia" topic and the event reporting the "Carlos Carrasco" case have a clear origin in journalists and public or health services. It also shows a uniformly distributed large audience, except for alternative therapy users (pseudo in Figure 3). At the other side of the figure, the "skin cancer" topic has more different types of authors, including the large number of tweets created by public and health services, probably as part of prevention campaigns. Finally, the rare case of the "tapeworm" has been reported by many authors (probably trying to clarify its meaning or to prevent its consequences) but retweeted by a small audience.

4.3. Controversial and Dicussion-Related Topics

Finally, to show the usefulness of the proposed analytical model, we have selected a group of events and topics whose parameters show that they are candidates for monitoring during PHS tasks. Table 8 shows some topics and events with a low user quality and relevant audience. Notice that, except for the first topic, none of them have verified accounts in their audience. Topics 1 and 2 take the part of campaigns. Topics 3 and 4 present very low Yule's I scores and a high number of messages in the same stream. These are clearly spam topics, used as a claim with non-informative purposes. Finally, Topic 5 is an example of an unverified event about some miraculous cure.

Table 8. Controversial events/topics.

Topic/Event	User's Quality (Y, %Ver., OnDom.)	#Users Twitting/Retw.	%Tweets (Replies, 1st_Person)
1. black skin	(85, 5%, 2)	180/5915	(43%, 22%)
2. force chemo	(97, 0%, 5)	66/15	(3%, 0%)
3. *evita metástasis*	(24, 0%, 282)	109/185	(0%, 0%)
4. *fármaco frenar*	(25, 0%, 255)	115/366	(0%, 0%)
5. *desaparece oración*	(56, 0%, 4)	83/3	(0%, 0%)

Regarding the nature of the tweets associated to these events/topics, the topic "black skin" associated to "skin cancer" has 43% of replies, indicating that is a very active discussion topic. The topic "force chemo" has only 3% of replies with few isolated opinions, which indicates a weak reaction against a government decision. The rest of topics/events in Table 8 do not contain any replies, despite there being many users writing about them.

4.4. Discussion

Nowadays, by applying modern Social Business Intelligence technologies, many organizations are developing solutions that integrate social media data into the multidimensional models of their decision-support systems. Here, one of the most important challenges is to adapt the analysis models to work with dynamic data streams and, in this way, to build systems that produce information in real time. The main contribution of the work presented in this paper is to apply this technology to the development of health surveillance systems. In this sense, the paper proposes a multidimensional model for Twitter streams analysis designed to exploit public health information in an intelligent way and demonstrates its usefulness for PHS tasks.

The proposed model annotates posts with attributes and metrics that describe their contents, users, quality and temporality. Then, these semantic annotations can be applied to filter and summarise the incoming data stream. The resulting system enables us to detect and track social media topics and events during PHS tasks. One strong point of the proposed method is the inclusion of quality metrics in the multidimensional model so that PHS analysis can be performed in a reliable way. Another strong point is the use of unsupervised statistical methods for identifying events, topics and profiling users. In this way, the multidimensional model can be adapted in real time to the data stream being monitored.

The use cases analysed in this paper demonstrate the advantages of the proposed tool for intelligent processing of social data and value creation in the context of PHS. With this conceptual framework, it is possible to derive a variety of dimensions to explain events and topics monitored by processes on demand. This is thanks to the linking and summarizing of the user-generated data (posts facts) to the users' actions and profiles (user facts). The proposed methodology allows health officers to address the main applications of social media in PHS, like social monitoring, situational awareness and communication surveillance.

The analysed use cases also show how to get insights into the most significant topics and events related to the tracked themes on Twitter, focusing on the nature and quality of the involved users. Analysing by author dimension and quality metrics allows us to elucidate the credibility of the source,

for example to determine if the information comes from official sources, human users or bots. Grouping posts metrics by frequent relevant topics facilitates a quick inspection of possible disease outbreaks as well as related events. For example, in our experiments the statistics of tweets and replies written in the first person grouped by the topic "leukemia" allowed us to locate self-reported experiences in therapy. On the other hand, event detection offers the possibility of estimating the current incidence of a disease. The same methodology could be used to detect self-reported humanitarian needs during disasters. Knowing the type of user that produces the information and its audience, and measuring their reactions to certain messages, makes it possible to understand how they relate to each other and to monitor the degree of awareness and perception of PHS events.

5. Conclusions

In this paper we have presented a novel method for the multidimensional analysis of Twitter streams aimed at helping PHS systems. The proposed method can automatically detect events and topics from the stream according to their semantics and temporal features, and to properly summarize their contents, as well as the users involved in publishing and interacting with them. Our experiments on a real long-term stream about cancer show the usefulness of these multidimensional models for PHS tasks. The proposed method places special emphasis on the quality of the data and the audience involved in the detected events and topics.

Future work will focus on defining new interesting key performance indicators (KPI) for PHS so that they can be estimated from the proposed multidimensional method in combination with other external data sources [42]. We also plan to improve the automatic classification of events and topics into simple taxonomies by using either external knowledge resources (e.g., UMLS Meta-thesaurus) and/or advanced text-mining techniques [30].

Author Contributions: Conceptualization, I.L., R.B. and M.J.A.; methodology, I.L and M.J.A.; software, R.B. and I.L.; validation, M.J.A., I.L. and R.B.; writing—original draft preparation, M.J.A.; writing—review and editing, R.B., M.J.A. and I.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Spanish Ministry of Industry and Commerce grant number TIN2017-88805-R and by the pre-doctoral grant of the Universitat Jaume I with reference PREDOC/2017/28.

Acknowledgments: We would like to thank the reviewers for their thoughtful comments and efforts towards improving our manuscript.

Conflicts of Interest: "The authors declare no conflict of interest."

References

1. Richards, C.L.; Lademarco, M.F.; Atkinson, D.; Pinner, R.W.; Yoon, P.; Mac Kenzie, W.R.; Frieden, T.R. Advances in Public Health Surveillance and Information Dissemination at the Centers for Disease Control and Prevention. *Public Health Rep.* **2017**, *132*, 403–410. [CrossRef]
2. Fung, I.C.; Tse, Z.T.; Fu, K.W. The use of social media in public health surveillance. *WPSAR* **2015**, *6*, 3–6. [CrossRef]
3. Jordan, S.E.; Hovet, S.E.; Fung, I.; Liang, H.; Fu, K.W.; Tse, Z. Using Twitter for Public Health Surveillance from Monitoring and Prediction to Public Response. *Data* **2018**, *4*, 6. [CrossRef]
4. Zhou, L.; Zhang, D.; Yang, C.; Wang, Y. Harnessing social media for health information management. *Electron. Commer. Res. Appl.* **2018**, *27*, 139–151. [CrossRef]
5. Al-Garadi, M.A.; Khan, M.S.; Varathan, K.D.; Mujtaba, G.; Al-Kabsi, A.M. Using online social networks to track a pandemic: A systematic review. *J. Biomed. Inform.* **2016**, *62*, 1–11. [CrossRef]
6. Adrover, C.; Bodnar, T.; Huang, Z.; Telenti, A.; Salathé, M. Identifying adverse effects of HIV drug treatment and associated sentiments using twitter. *JMIR Public Health Surveill* **2015**, *1*, 7. [CrossRef]
7. Prieto, V.M.; Matos, S.; Alvarez, M.; Cacheda, F.; Oliveira, J.L. Twitter: A good place to detect health conditions. *PLoS ONE* **2014**, *9*, e86191. [CrossRef]

8. Ginn, R.; Pimpalkhute, P.; Nikfarjam, A.; Patki, A.; O'Connor, K.; Sarker, A.; Gonzalez, G. Mining Twitter for adverse drug reaction mentions: A corpus and classification benchmark. In Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing, Reykjavík, Iceland, 27 May 2014.
9. Sarker, A.; Gonzalez, G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *J. Biomed. Inform.* **2015**, *53*, 196–207. [CrossRef]
10. Yepes, A.J.; MacKinlay, A.; Han, B. Investigating public health surveillance using Twitter. *ACL-IJCNLP* **2015**, *15*, 164.
11. Myslín, M.; Zhu, S.-H.; Chapman, W.; Conway, M. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *J. Med. Int. Res.* **2013**, *15*, 174. [CrossRef]
12. Ji, X.S.; Chun, A.; Wei, Z.; Geller, J. Twitter sentiment classification for measuring public health concerns. *Soc. Netw. Anal. Min.* **2015**, *5*, 13. [CrossRef]
13. Santos, J.C.; Matos, S. Analysing Twitter and web queries for flu trend prediction. *Theor. Biol. Med. Model.* **2014**, *11*, S6. [CrossRef]
14. Paul, M.J.; Dredze, M. A model for mining public health topics from Twitter. *Health* **2012**, *11*, 16.
15. Missier, P.; Romanovsky, A.; Miu, T.; Pal, A.; Daniilakis, M.; Garcia, A.; da Silva Sousa, L. Tracking dengue epidemics using twitter content classification and topic modelling. In Proceedings of the 16th International Conference onWeb Engineering, Lugano, Switzerland, 6–9 June 2016; pp. 80–92.
16. Prier, K.W.; Smith, M.S.; Giraud-Carrier, C.; Hanson, C.L. Identifying health-related topics on twitter. In Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, College Park, MD, USA, 30–31 March 2011; pp. 18–25.
17. Pennacchiotti, M.; Popescu, A.M. A Machine Learning Approach to Twitter User Classification. In Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Spain, 17–21 July 2011; pp. 281–288.
18. Zaeem, R.N.; Liau, D.; Barber, K.S. Predicting Disease Outbreaks Using Social Media: Finding Trustworthy Users. In Proceedings of the Future Technologies Conference (FTC) 2018. FTC 2018. Advances in Intelligent Systems and Computing, Vancouver, BC, Canada, 13–14 November 2018; Arai, K., Bhatia, R., Kapoor, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2019; Volume 880.
19. Bian, J.; Zhao, Y.; Salloum, R.G.; Guo, Y.; Wang, M.; Prosperi, M.; Sun, Y. Using Social Media Data to Understand the Impact of Promotional Information on Laypeople's Discussions: A Case Study of Lynch Syndrome. *J. Med. Internet Res.* **2017**, *19*, e414. [CrossRef]
20. Gomide, J.; Veloso, A.; Meira, W.; Almeida, V.; Benevenuto, F.; Ferraz, F.; Teixeira, M. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. In Proceedings of the 3rd International Web Science Conference, Koblenz, Germany, 15–17 June 2011; p. 3.
21. Denecke, K.; Harries, M.; Otrusina, L.; Smrz, P.; Dolog, P.; Nejdl, W.; Velasco, E. How to Exploit Twitter for Public Health Monitoring? *Methods Inf. Med.* **2013**, *50*, 326–339.
22. Zadeh, A.; Zolbanin, H.; Sharda, R.; Delen, D. Social Media for Nowcasting Flu Activity: Spatio-Temporal Big Data Analysis. *Inf. Syst. Front.* **2019**, *21*, 743–760. [CrossRef]
23. Dredze, M.; Paul, M.J.; Bergsma, S.; Tran, H. Carmen: A Twitter geolocation system with applications to public health. In Proceedings of the AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI), Bellevue, DC, USA, 14–18 July 2013; pp. 20–24.
24. Liang, H.; Shen, F.; Fu, K.W. Privacy protection and self-disclosure across societies: A study of global Twitter users. *New Media Soc.* **2017**, *19*, 1476–1497. [CrossRef]
25. Inmon, W. *Building the Data Warehouse*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2005.
26. Abelló, A.; Romero, O.; Bach Pedersen, T.; Berlanga, R.; Nebot, V.; Aramburu, M.J.; Simitsis, A. Using Semantic Web Technologies for Exploratory OLAP: A Survey. *IEEE Trans. Knowl. Data Eng.* **2015**, *27*, 571–588. [CrossRef]
27. Akter, S.; Bhattacharya, M.; Fosso Wamba, S.; Aditya, S. How does Social Media Analytics Create Value? *J. Organ. End User Comput.* **2016**, *28*, 1–9. [CrossRef]
28. Berlanga, R.; Jiménez-Ruiz, E.; Nebot, V. Exploring and linking biomedical resources through multidimensional semantic spaces. *BMC Bioinform.* **2012**, *13*, e17. [CrossRef]
29. Lanza-Cruz, I.; Berlanga, R.; Aramburu, M.J. Modeling Analytical Streams for Social Business Intelligence. *Informatics* **2018**, *5*, 33. [CrossRef]

30. Berlanga, R.; Lanza-Cruz, I.; Aramburu, M.J. Quality Indicators for Social Business Intelligence. In Proceedings of the 6th International Conference on Social Network Analysis, Management & Security (SNAMS 2019), Granada, Spain, 22–25 October 2019.
31. Kim, Y.; Huang, J.; Emery, S. Garbage in, Garbage Out: Data Collection, Quality Assessment and Reporting Standards for Social Media Data Use in Health Research, Infodemiology and Digital Disease Detection. *J. Med. Internet Res.* **2016**, *18*, e41. [CrossRef]
32. Allem, J.P.; Ferrara, E. The Importance of Debiasing Social Media Data to Better Understand E-Cigarette-Related Attitudes and Behaviors. *J. Med. Internet Res.* **2016**, *18*, e219. [CrossRef]
33. Massoudi, K.; Tsagkias, M.; de Rijke, M.; Weerkamp, W. Incorporating query expansion and quality indicators in searching microblog posts. In *European Conference on Information Retrieval*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 18–21.
34. Xie, W.; Zhu, F.; Jiang Lim, P.; Wang, K. TopicSketch: Real-Time Bursty Topic Detection from Twitter. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 2216–2229. [CrossRef]
35. Momeni, E.; Tao, K.; Haslhofer, B. Identification of Useful User Comments in Social Media: A Case Study on Flickr Commons. In Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, Indianapolis, IN, USA, 22–26 July 2013; pp. 1–10.
36. Chen, W.; Yeo, C.K.; Lau, C.T.; Lee, B.S. A study on real-time low-quality content detection on Twitter from the users' perspective. *PLoS ONE* **2017**, *12*, 8. [CrossRef]
37. Feng, W.; Zhang, C.; Zhang, W.; Han, J.; Wang, J.; Aggarwal, C.; Huang, J. STREAMCUBE: Hierarchical spatio-temporal hashtag clustering for event exploration over the Twitter stream. In Proceedings of the 2015 IEEE 31st International Conference on Data Engineering, Seoul, Korea, April 13–16 2015; pp. 13–17.
38. Zhou, X.; Chen, L. Event detection over twitter social media streams. *VLDB J.* **2014**, *23*, 381–400. [CrossRef]
39. Zubiaga, A.; Spina, D.; Martínez, R.; Fresno, V. Real-time classification of Twitter trends. *J. Assoc. Inf. Sci. Technol.* **2015**, *66*, 462–473. [CrossRef]
40. Berlanga, R.; García-Moya, L.; Nebot, V.; Aramburu, M.J.; Sanz, I.; Llidó, D. SLOD-BI: An Open Data Infrastructure for Enabling Social Business Intelligence. *Int. J. Data Warehous. Data Min.* **2015**, *11*, 1–28. [CrossRef]
41. Liu, X.; Tang, K.; Hancock, J.; Han, J.; Song, M.; Xu, R.; Pokorny, B. A Text Cube Approach to Human, Social and Cultural Behavior in the Twitter Stream. In Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, Washington, DC, USA, 2–5 April 2013.
42. Rosenthal, S.; Farra, N.; Nakov, P. Sentiment Analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017.
43. Rodríguez-Vidal, J.; Gonzalo, J.; Plaza, L.; Anaya, H. Automatic detection of influencers in social networks: Authority versus domain signals. *J. Assoc. Inf. Sci. Technol.* **2019**, *70*, 675–684. [CrossRef]
44. Mahalakshmi, G.S.; Koquilamballe, K.; Sendhilkumar, S. Influential Detection in Twitter Using Tweet Quality Analysis. In Proceedings of the Second International Conference on Recent Trends and Challenges in Computational Models, Tindivanam, India, 3–4 February 2017; pp. 315–319.
45. Miller, Z.; Dickinson, B.; Deitrick, W.; Hu, W.; Wang, A.H. Twitter spammer detection using data stream clustering. *Inf. Sci.* **2014**, *260*, 64–73. [CrossRef]
46. Varol, O.; Ferrara, E.; Davis, C.A.; Menczer, F.; Flammini, A. Online Human-Bot Interactions: Detection, Estimation, and Characterization. In Proceedings of the Eleventh International AAAI Conference on Web and Social Media, Montreal, QC, Canada, 15–18 May 2017.
47. Li, H.; Mukherjee, A.; Liu, B.; Kornfield, R.; Emery, S. Detecting Campaign Promoters on Twitter using Markov Random Fields. In Proceedings of the IEEE International Conference on Data Mining, Shenzhen, China, 4–5 March 2014.
48. Francia, M.; Gallinucci, E.; Golfarelli, M.; Rizzi, S. Social Business Intelligence in Action. In Proceedings of the Advanced Information Systems Engineering 28th International Conference CAiSE, Ljubljana, Slovenia, 13–17 June 2016; pp. 33–48.
49. Liu, B. *Sentiment Analysis and Opinion Mining*; Morgan & Claypool Publishers: San Rafael, CA, USA, 2012.
50. García-Moya, L.; Anaya-Sánchez, H.; Berlanga, R. A Language Model Approach for Retrieving Product Features and Opinions from Customer Reviews. *IEEE Intell. Syst.* **2013**, *28*, 19–27. [CrossRef]
51. Guille, A.; Hacid, H.; Favre, C.; Zighed, D.A. Information Diffusion in Online Social Networks: A Survey. *SIGMOD Record* **2013**, *42*, 17–28. [CrossRef]

52. Nadal, S.; Herrero, V.; Romero, O.; Abelló, A.; Franch, X.; Vansummeren, S.; Valerio, D. A software reference architecture for semantic-aware Big Data systems. *Inf. Softw. Technol.* **2017**, *90*, 75–92. [CrossRef]
53. Javed, M.H.; Lu, X.; Panda, D.K. Characterization of Big Data Stream Processing Pipeline: A Case Study using Flink, Kafka. In Proceedings of the Fourth IEEE/ACM International Conference on Big Data Computing, Applications, Technologies, New York, NY, USA, 5–8 December 2017; pp. 1–10.
54. Okugami, C.; Sparks, R.; Woolford, S. Twitter Data Offers Opportunities for Public Health Professionals. *J. Health Med. Inf.* **2014**, *5*, e123. [CrossRef]
55. Xu, S.; Markson, C.; Costello, K.L.; Xing, C.Y.; Demissie, K.; Llanos, A. Leveraging Social Media to Promote Public Health Knowledge: Example of Cancer Awareness via Twitter. *JMIR Public Health Surveill.* **2016**, *2*, e17. [CrossRef] [PubMed]
56. Yules, C.U. *The Statistical Study of Literary Vocabulary*; Cambridge Press: Cambridge, UK, 1944.
57. Clauset, A.; Newman, M.E.; Moore, C. Finding community structure in very large networks. *Physical Review. E* **2004**, *70*, 066111. [CrossRef]

Chapter 7

Final remarks

7.1 General Discussion

The field of Social Business Intelligence (SBI) faces the challenge of capturing strategic information from social networks. Unlike traditional Business Intelligence (BI), SBI must deal with the high dynamics of both social media content and the analytical demands of the enterprise, in addition to the vast amount of noisy data. To effectively harness these continuous sources of data, efficient streaming data processing is required to give them semantic structure and transform them into enlightening facts. In this context, the developed chapters in this thesis addressed the challenge of the efficient developing of social indicators. In this section, the results of each chapter are examined, along with their correspondence to the related objectives and hypotheses (see sections 1.5 and 1.6). This will help understand how these findings are integrated into the complex task of implementing SBI.

The following chapters address goal 1 and its related hypothesis 1:

- **Chapter 2:** "Defining Dynamic Indicators for Social Network Analysis: A Case Study in the Automotive Domain using Twitter."

  This chapter addresses a methodological framework for specifying and monitoring key indicators derived from real-time social media metrics. The framework is presented as a comprehensive solution for collecting, analysing, and

understanding these real-time metrics, providing professionals and organizations with the ability to make informed and strategic decisions in the dynamic environment of social networks.

The key components of the methodological framework are as follows: (i) Specification of metrics, indicators, and dimensions of interest. An ontological model is described to define social indicators. The article discusses the modelling of social data as a multidimensional cube, which facilitates analysis and visualization from various perspectives, as required in the objective. (ii) Data collection, where it explains how the necessary data is collected and organized to calculate the indicators continuously and in real-time. (iii) Collection and automated data processing. The developed methods include data extraction, transformation, and loading processes over linked data models. Virtual and dynamic data cubes are generated based on defined analytical patterns. The article's focus is on real-time data stream processing, which is essential for monitoring and making decisions based on constantly evolving social indicators. (iv) Finally, the component related to monitoring and visualization explains how the observations of the defined indicators are calculated and presented.

It's worth noting that the involved data and information are represented and stored as Linked Data, as required by the objective. The implementation of semantic technologies allows for a more precise and contextual representation of the data, facilitating advanced analysis and more informed decision-making. The proposed framework constitutes a significant contribution in the field of SBI (Social Business Intelligence), from which an organization can enhance decision-making processes, detect real-time trends, or improve understanding of social metrics. The chapter provides evidence of how the implementation of a framework based on semantic models and real-time Twitter data processing methods contributes to the identification of social patterns in the automotive domain. It is demonstrated that these approaches can be successfully used to analyse and understand user behavior on social networks.

- **Chapter 3:** "Modelling Analytical Streams for Social Business Intelligence."

In this chapter, a methodological framework is introduced, which represents a substantial improvement over the initial framework proposed in Chapter 1. The new framework suggests a multidimensional formalism for representing and evaluating social indicators directly from event streams derived from social network data. This formalism is based on two main aspects: the semantic representation of events through Linked Data and the support of multidimensional analysis models similar to OLAP. Based on the specifications of the analysis objectives, all the necessary event streams are modeled and deployed to track the indicators. A Kappa-type streaming architecture is proposed for this purpose.

The key differences and benefits of the new proposal are as follows: (i) A new approach to continuous analysis and enrichment of data streams, enabling real-time processing of social network information. A workflow model is proposed for processing data streams, based on which raw data streams can be transformed into three types of analytical data streams (Data, RDF, and Fact streams) to feed different analytical tasks and machine learning models. The use of Linked Data and multidimensional modelling facilitates data enrichment and workflow validation,

increasing the reliability of results. (ii) Addressing dynamic challenges: The new proposal tackles specific challenges associated with real-time analysis of social data, including the dynamics of elements involved in the analysis and the need for intelligent processing for analytical tasks such as sentiment analysis and spam detection. Thus, the framework is better suited to address the specific challenges of Social Business Intelligence (SBI). (iii) Adoption of a streaming architecture that establishes channels and dynamics for information exchange between the system and its actors. (iv) The solution aims to integrate Data Science and Data Analysis tasks into a single working environment, which is essential for addressing complex challenges in real-time social data analysis.

This research further reinforces the initial hypothesis by proposing an architecture that combines semantic data representation with dynamic multidimensional analysis in real-time. This article also provides a use case example in the automotive sector, suggesting that this methodology is applicable in various domains. Given the research conducted, the presented results are consistent and reinforce the idea that the implementation of frameworks based on semantic models and real-time data processing methods enhances the identification and tracking of social patterns. It is reasonable to conclude that the research hypothesis is positively supported.

The following chapter address goal 2 and its related hypothesis 2:

- **Chapter 4:**"Multidimensional Author Profiling for Social Business Intelligence."

The research introduces an innovative author profiling method specifically aimed at classifying social media users in multidimensional perspectives for Social Business Intelligence (SBI) applications. This approach is relevant because it adapts to the dynamics of SBI, where user profiles are defined according to the specific needs of each SBI application. The method is unsupervised and allows for obtaining labelled datasets necessary for training profile classifiers. The approach is based on the use of user descriptions, which are part of post metadata, rather than analysing the content of tweets, making more efficient use of data storage and processing time resources. It's worth noting that the experimental results achieve promising performance in user profile classification. Finally, empirical results in different domains demonstrate the usefulness and effectiveness of the proposed method.

The results provide evidence that it is possible to develop an author profiling method in social networks that aligns with the multidimensional and dynamic perspectives required in SBI projects. This supports the thesis-related hypothesis and constitutes a significant contribution to both the scientific community and the business sector.

The following chapter address goal 3 and its related hypothesis 3:

- **Chapter 5:** " Quality Management in Social Business Intelligence Projects."

The implementation of a business intelligence system should commence with the establishment of mechanisms to ensure the construction of high-quality data collections. This stage is crucial for developing robust models and information systems capable of providing reliable results for analysis and decision-making. Indeed, the validity of any information system largely depends on data quality. However, it is worth noting that, within the realm of Social Business Intelligence, data quality management has often been underestimated. Most approaches rely on building data collections using simple keyword-based filters or ad-hoc rules, leading to the selection of noisy and low-quality data.

Within the context of this thesis, there was a need to develop an efficient infrastructure for managing social indicators, and to address this challenge, the third objective was established. Chapter 5 offers solutions to overcome the limitations identified in the literature, presenting innovative methodological approaches for constructing high-quality social data collections. The two research articles that constitute a sequence of improvements are: "Quality Indicators for Social Business Intelligence" and "A Data Quality Multidimensional Model for Social Media Analysis."

Our primary hypothesis is based on the premise that data quality is significantly influenced by both the relevance of the users generating it and the coherence of their posts in relation to the application domain. Consequently, the development of the solution consistently aligns with the chosen theories and premises.

In summary, the primary contribution of this chapter lies in the introduction of a multidimensional framework for dynamically and semi-automatically defining quality indicators to assess social media data collections. The proposed multidimensional analysis methodology offers the advantage of evaluating the quality of social data from various multidimensional perspectives, adapted to dynamic analysis contexts, such as dimensions of interest, domain-specific language models, and analysis objectives. It also allows for different levels of granularity of defined analytical patterns, ranging from individual posts to entire data collections.

It is demonstrated that user profiling based on their business role is a key quality indicator for implementing the proposed approach in quality indicator discovery. The process of labeling relevant users can be automated using the author profiling method presented in Chapter 4. Another practical approach to determine the value of metrics from a business role perspective would be to apply the author profiling method to all data stream users and consider as relevant only those whose classification matches the corresponding user role perspective.

Finally, the proposed method offers vital insights for assessing analytical data quality and making decisions about data filtering and parameter updates. This may lead analysts to reconsider certain extraction process parameters, like the keywords and reference user set. Subsequent iterations can be performed to improve the dataset, guided by automatically generated quality indicators that assess the impact of these actions on the results. The practical results provide evidence that the hypothesis and objectives were successfully met.

The following chapter address goal 4:

- **Chapter 6:** "Social Media Multidimensional Analysis for Intelligent Health Surveillance."

The chapter presents an innovative approach to the integration of social media data into decision support systems, particularly in the context of public health surveillance. This approach represents the final phase of research aimed at establishing an evaluation framework that incorporates the methods proposed in the thesis to manage and analyse strategic predictive social indicators.

First and foremost, the chapter focuses on data quality by proposing a multidimensional approach that includes quality metrics and filtering techniques. The quality metrics enable the removal of irrelevant users, thereby avoiding the inclusion of non-pertinent information that could bias results and conclusions. Furthermore, the ability to evaluate the credibility of users as sources of information and to identify relevant profiles is highlighted. In this way, data quality is assessed and improved, emphasizing users who possess valuable information related to public health.

The multidimensional model proves effective in detecting relevant events and topics in Twitter streams, as well as in analysing their audience and real-time impact. This underscores the effectiveness of the indicators in identifying events and topics of interest in the context of analysis. The model also facilitates the assessment of the audience and impact of identified events and topics, which is essential for understanding how predictions can affect the community and support decision-making. Furthermore, the use of unsupervised statistical methods to identify the main analytical patterns in real-time is emphasized. This feature allows the model to dynamically adapt to continuously changing data, which is crucial in public health surveillance, where situations can evolve rapidly.

In summary, the multidimensional approach to managing data quality, defining social indicators, the effectiveness of these indicators, and the reliability of predictive methods strongly support the thesis's objective of establishing an evaluation framework that addresses these essential aspects in the context of strategic decision-making in public health.

7.2 Main Contributions

This PhD thesis addresses the problem of defining a methodological framework for the definition and monitoring of novel social indicators for SBI. To achieve this, it was necessary to integrate several established methodologies and techniques while introducing innovative approaches to tackle the specific challenges of dynamic social indicators. The main contributions of this thesis can be summarized as follows:

1. Development of a comprehensive literature review. A thorough review of existing literature on social indicators and SBI methodologies was conducted. This review provided a solid theoretical foundation where gaps and opportunities for novel contributions are identified.

2. Development of a semantic model to represent the data analytic patterns of social networks relevant for the development of meaningful dynamic social indicators for SBI. The model follows the LD principles and fits linked data into a multidimensional model. The model enables the definition of multidimensional structures related to the context of analysis, domain, topics, and business categories for user segmentation. Additionally, it allows the representation of strategic social indicators and potential links with external linked data sources. Moreover, it helps to discover new valuable information through statistical analysis or predictive techniques over the data. The aforementioned principles are embodied in the proposal of an ontology to represent and monitor dynamic social indicators, as well as their relationships with the different elements of the business model. One main benefit is the fact that indicators are directly linked to the social measures, so that is possible to easily identify the origin the values of these indicators. On the other hand, as indicators are semantic data, it is possible to apply validation techniques during their definition and derivation. This formal representation constitutes, to our knowledge, the first proposal to represent and organize social indicators as a knowledge graph.

3. A multidimensional framework is proposed to represent and evaluate social indicators directly from fact-based streams derived from social network data. Specifically, data streams are processed based on the semantic representation of facts through LD and the support of OLAP-like multidimensional analysis models. These models facilitate the semantic enrichment of the data, and their representation according to analysis requirements. The approach allows the on-demand definition of social indicators, as well as the dynamic handling of dimensions and metrics. Additionally, a Kappa-like architecture is proposed for streaming processing of social indicators. It enables the integration of both Data Science and Data Analysis tasks within the same workspace. In summary, this contribution proposes a comprehensive and robust approach to represent, evaluate, and analyse dynamic social indicators from data streams derived from social networks, providing precision, semantic enrichment, flexibility, and integration capability in the analysis.

4. A novel multidimensional author profiling method for SBI is proposed. The proposed method involves analysing and classifying social network users based on multidimensional perspectives of user business roles. For this purpose, the method utilizes linguistic patterns and semantic features extracted from their profile descriptions. Semantic features are linked to semantic knowledge represented by ontologies. This allows us to verify consistency and to identify conflicts in labels within the training data. By applying a multidimensional perspective, the method takes into account multiple dimensions or aspects of author profiles, allowing for a more comprehensive and accurate analysis. The main novelty of the approach is that it allows the unsupervised construction of a labelled dataset that serves as input for text classification tasks. In addition, the process could be adapted to dynamic scenarios, an important requirement in SBI projects. Social media data is constantly evolving and changing, and the ability to handle dynamic scenarios allows the method to effectively analyse and classify authors even when faced with new or evolving linguistic patterns and trends in social media content. These advancements contribute to more accurate and efficient analysis of authors in social media.

5. We propose a method that addresses the crucial task of creating high-quality data collections specifically designed for social network analysis. This method applies a novel multidimensional data model for constructing cubes with impact measures for various quality metrics. Additionally, the quality cubes include a user-role dimension so that quality metrics can be evaluated from different user's perspectives. This data model enables a comprehensive and structured representation of data, facilitating the analysis of different quality aspects. Furthermore, the method enables the automatic construction of quality indicators tailored to analysis needs. This saves time and effort by automating the process, and it ensures that the quality indicators align with the specific requirements of the analysis tasks. The main practical implication of this method is that it allows analysts to measure the quality of processed data from different perspectives. Considering the categories of user profiles that interact with the contents the analysis incorporates a more holistic view, leading to a comprehensive understanding of the data quality. This contributes to improved decision-making and ensures that the data used for social network analysis is reliable and accurate.

6. The proposed framework was applied to real-world scenarios across several domains, including automotive, tourism, and public health. It underwent rigorous evaluation and validation through extensive case studies and comparative analyses. The developed indicators framework demonstrates the ability to identify relevant events and topics within the corresponding analysis domain. The developed methods offers valuable insights and real-time information on significant events and emerging trends. By incorporating semantics and temporal patterns, the approach enhances the accuracy and contextual understanding of social media phenomena.

7. Contribution to the field: This thesis makes a significant contribution to the field of SBI by providing a comprehensive methodological framework that streamlines the definition and monitoring of dynamic social indicators. It

effectively addresses significant gaps and limitations identified in the existing literature, providing valuable insights and practical guidance for researchers and practitioners working in this field.

7.3 Final Conclusions and Future Work

In the academic and business domains, social media data analytics have demonstrated their value in enhancing business intelligence strategies. However, their effective implementation requires a strategic approach and thorough analysis. Existing proprietary software tools for social analytics primarily concentrate on standard social metrics, overlooking the contextual information, crucial analytical patterns, and potential data relationships embedded within social data. As a result, this valuable information remains hidden and inaccessible. Academic proposals frequently focus on solving problems within closed contexts, often neglecting the dynamic nature of analysis needs. There is an increasing need for standardized methodologies in social network analytics to optimize their impact on strategic decision-making. To bridge this gap, this thesis presents a methodological framework for specifying and monitoring social indicators, encompassing comprehensive methods and techniques for Social Business Intelligence (SBI) projects.

In a broader context, the findings of this doctoral thesis have confirmed the viability and effectiveness of the proposed methodological framework for specifying and utilizing dynamic social indicators. The study has highlighted the significance of semantic technologies in comprehending social data and enabling information systems interoperability. Notably, during the validation process of the proposed methods, potential practical applications have been identified across economically significant sectors, such as automotive, tourism, and public health. This not only demonstrates the versatility of the proposal but also emphasizes its potential for adaptable implementation across different domains. It is essential to emphasize that the proposed system is dynamic, providing mechanisms to adjust and respond to changes and challenges in the business environment. Furthermore, it enables the collection, analysis, and presentation of real-time relevant information, establishing a strong foundation for strategic and operational decision-making.

The research findings have revealed new tasks and areas that lead to further exploration, offering potential directions for future work to expand upon the insights discovered in this study.

The first task is to automate the semantic definition of the indicators, as well as the queries necessary to retrieve the social metrics that feed them. It is also necessary to develop new predictive algorithms to determine the most appropriate metrics for certain strategic objective. We are considering implementing an improved BI model to derive indicators based on the way they relate to each other and to strategic business objectives. Specifically, as future work, we want to implement an approach to predict the best combination of metrics and indicators to generate composite indicators (index type) so that they best reflect the strategic objective to be measured. To achieve this, use could be made of features such as hidden patterns in the network relationships between the indicators and the strategic objectives associated with them within the knowledge graph, as well as the textual and semantic information related to the metadata. In this sense, we have undertaken a first approach to automatically align indicators with sustainable development goals

(SDGs) (paper 5 in section 7.4.1.2). Another crucial aspect to be addressed is the automatic recommendation of the most suitable formulas for combining social metrics, to effectively represent a specific trend or phenomenon. This area of research holds the potential to enhance our understanding of the optimal ways to aggregate and interpret social data, thereby enabling more accurate and insightful analysis. Future research endeavours could focus on developing intelligent algorithms that dynamically recommend the most appropriate approach for combining social metrics, taking into account the specific context and objectives of the analysis.

In relation to the task of profiling authors, as future work we consider improving the predictive models through the study of new analytical patterns, such as social metrics and semantic relationships between data, and the application of techniques applied to graphs of the generated knowledge. This strategy will also make it possible to classify users for whom there is no textual description. We plan to conduct a contrasting study to evaluate how well the user descriptions match the content they generate, and to find correlations between the descriptions and other aspects such as psychological traits and emotions of the users.

One of the main limitations we face during the development of a software project is the bias in the data, which can be present both in the data sources and in the formulation of the predictive models themselves. As future work, it is necessary to study and evaluate the best methods for bias correction in order to guarantee the variability of social data during the training of predictive models and thus obtain more accurate decisions.

With regard to the task of quality management in an SBI project, future work will mainly aim at addressing the main limitations mentioned in the related articles of this thesis. We must emphasize the need to study new methods that allow combining techniques such as semantic annotations and sentence encoders with the aim of improving the language models used in our methodology to measure some quality metrics, like the coherence, usefulness and completeness. Further research is also need on finding out new ways of combining quality metrics in order to maximise their utility. Additionally, it is crucial to evaluate whether the metrics obtained during the quality management process can also be utilized as input for enhancing the process itself. Specifically, exploring the potential of incorporating these metrics into the selection of relevant users could lead to the identification of new and more valuable quality indicators. This investigation would involve analysing the feedback loop between the quality management process and the metrics collected, aiming at optimizing the selection criteria and further enhance the accuracy and relevance of the obtained indicators.

7.4 Derived Research

In this section, the scientific publications derived from this thesis are listed. Additionally, the research projects developed based on the methodological framework for the development of dynamic social indicators proposed in this thesis are also included.

7.4.1 Scientific Publications

7.4.1.1 Journal Papers

1. Lanza Cruz, Indira; Berlanga, Rafael and Aramburu, María José. (2023). Multidimensional Author Profiling for Social Business Intelligence. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-023-10370-0. (Q1)
2. Aramburu, María José; Berlanga, Rafael and Lanza Cruz, Indira. (2023). A Data Quality Multidimensional Model for Social Media Analysis. *Business & Information Systems Engineering*, DOI: 10.1007/s12599-023-00840-9 (In Press). (Q1)
3. Aramburu, María José; Berlanga, Rafael and Lanza Cruz, Indira. (2020). Social Media Multidimensional Analysis for Intelligent Health Surveillance. *International Journal of Environmental Research and Public Health*, 17, 2289. https://doi.org/10.3390/ijerph17072289. (Q2)
4. Berlanga, Rafael; Aramburu, María José; Lanza Cruz, Indira; Llidó Escrivá, D.M.; Museros, Lledó and Sanz, Ismael. (2018). Dynamic SLOD-BI: Infraestructura Dinámica de Inteligencia de Negocio Social. *Actas de las XXIII Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2018)*. SISTEDES. http://hdl.handle.net/11705/JISBD/2018/045.
5. Lanza Cruz, Indira; Berlanga, Rafael and Aramburu, María José. (2018). Modeling Analytical Streams for Social Business Intelligence. *Informatics*, 5, 33. https://doi.org/10.3390/informatics5030033. (Q1)
6. Lanza Cruz, Indira, Aramburu, María José and Berlanga, Rafael. (2016). Metodología de inteligencia de negocio para análisis social en la infraestructura de datos enlazados slod-bi. *Ciência da Informação*, v.45, n.3, 2016. DOI: 10.18225/ci.inf..v45i3.4058

7.4.1.2 Conference papers

1. Aramburu, María José; Berlanga, Rafael and Lanza Cruz, Indira. (2021). Quality Management in Social Business Intelligence Projects. *In Proceedings of the 23rd International Conference on Enterprise Information Systems - Volume 1: ICEIS*, ISBN 978-989-758-509-8; ISSN 2184-4992, pages 320-327. DOI: 10.5220/0010495703200327.
2. Berlanga, Rafael; Lanza Cruz, Indira and Aramburu, María José. (2019). Quality Indicators for Social Business Intelligence. *In 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Granada, Spain, pp. 229-236, doi: 10.1109/SNAMS.2019.8931862.
3. Lanza Cruz, Indira; Berlanga, Rafael. (2018). Defining Dynamic Indicators for Social Network Analysis: A Case Study in the Automotive Domain using Twitter. *In Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018) – KDIR*. ISBN 978-989-758-330-8; ISSN 2184-3228, SciTePress, pages 221-228. DOI: 10.5220/0006932902210228.
4. Berlanga, Rafael; Yepes, A.J.; Pérez, M. and Lanza Cruz, Indira. (2018). Coarse-grained Semantic Characterization of Large Knowledge Resources. *In Proceedings of the 5th Spanish Conference on Information Retrieval (CERI '18)*. Association for Computing Machinery, New York, NY, USA, Article 16, 1–4. https://doi.org/10.1145/3230599.3230616.
5. Soriano, Mario; Berlanga, Rafael and Lanza Cruz, Indira. (2023). On the problem of automatically aligning indicators to SDGs. *In: Pesquita, C., et al. The Semantic Web: ESWC 2023 Satellite Events. ESWC 2023. Lecture Notes in Computer Science*, vol 13998. Springer, Cham. https://doi.org/10.1007/978-3-031-43458-7_26.

7.4.2 Related Projects and Research Actions

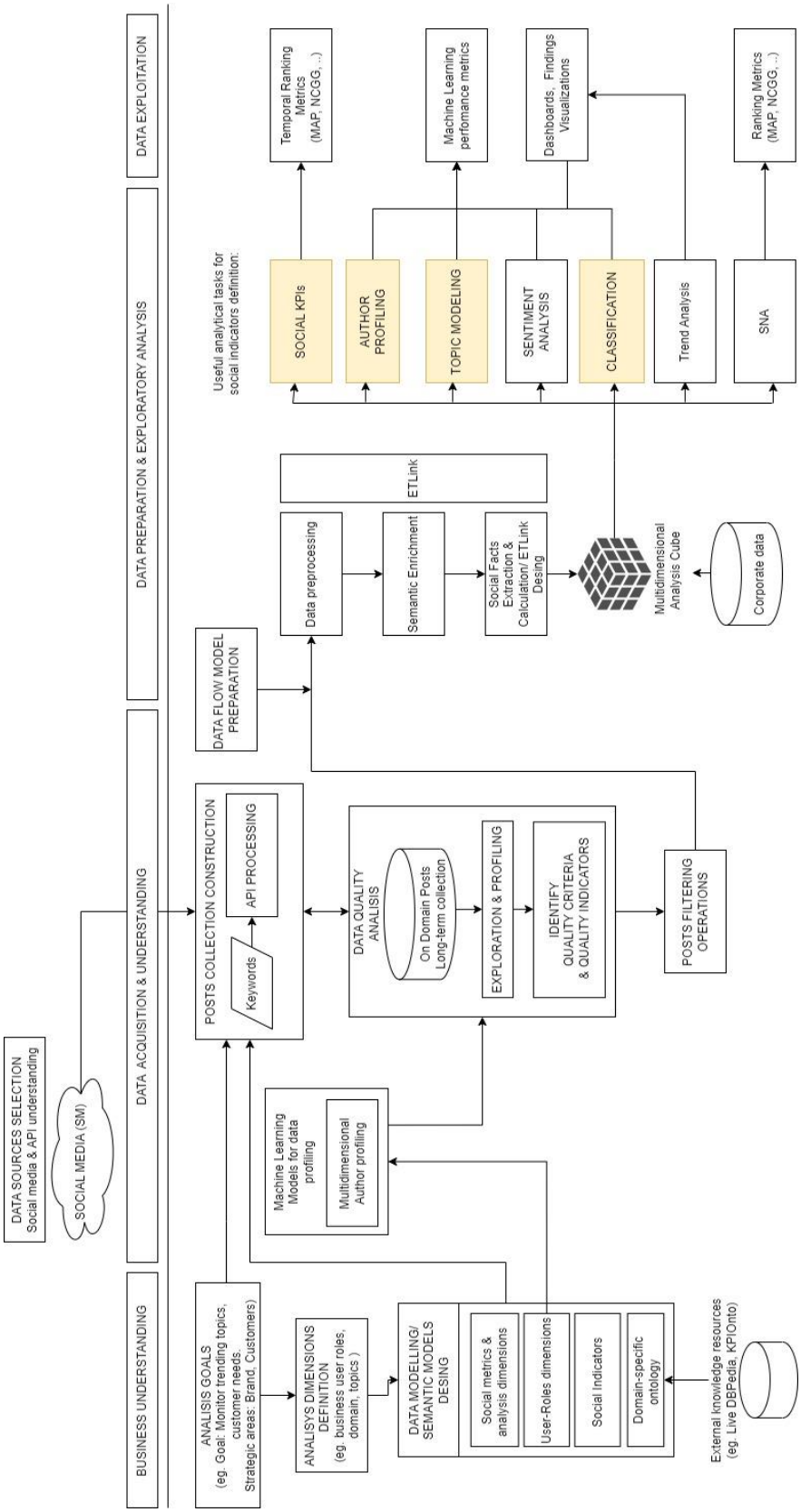
1. Title: PRUEBA DE CONCEPTO PARA LA PLATAFORMA DE ANÁLISIS SOCIAL DINÁMICO EN EL CONTEXTO DEL TURISMO SOSTENIBLE
   Reference: PDC2021-121097-I00
   Lead researcher: Rafael Berlanga Llavori
   Awarded funding: 69,000.00€
   Funding organization: AGENCIA ESTATAL D'INVESTIGACIÓ
   Start date: 01-12-2021 End date: 30-11-2023

2. Title: APLICACIÓN DE TÉCNICAS DE IA PARA LA GENERACIÓN DE DIGITAL TWINS EN TURISMO
   Reference: UJI-B2020-15
   Lead researcher 1: Lledó Museros Cabedo
   Lead researcher 2: Rafael Berlanga Llavori
   Awarded funding: 15,966.30€
   Funding organization: UNIVERSITAT JAUME I
   Start date: 01-01-2021 End date: 31-12-2023

3. Title: INFRAESTRUCTURA DINÁMICA DE INTELIGENCIA DE NEGOCIO SOCIAL PARA EMPRESAS PYME
   Reference: TIN2017-88805-R
   Lead researcher: Rafael Berlanga Llavori
   Awarded funding: 57,717.00€
   Funding organization: MINISTERIO DE ECONOMÍA, INDUSTRIA Y COMPETITIVIDAD
   Start date: 01-01-2018 End date: 30-09-2021

4. Title: STUDY AND DEVELOPMENT OF A METHODOLOGY FOR TOPIC IDENTIFICATION ACCORDING TO THE DOMAIN CONTEXT WITHIN THE SOCIAL CONVERSATION.
   Reference: E-2019-16
   Code: 19I187.16/1

Appendix A



Appendix A. Methodological Framework