A Computational Academic Integrity Framework

Ph.D. Candidate: Alexander Amigud
Supervisor: Dr. Thanasis Daradoumis
Co-supervisor: Dr. Joan Arnedo-Moreno

A thesis submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Network and Information Technologies

Department of Computer Science, Multimedia and Telecommunications
Universitat Oberta de Catalunya (UOC) Barcelona, Spain.

January 16, 2018

To the creator of Homo Analyticus— the ultimate machine learning machine.

Foreword

And when he died, I suddenly realized I wasn't crying for him at all, but for the things he did. I cried because he would never do them again, he would never carve another piece of wood or help us raise doves and pigeons in the backyard or play the violin the way he did, or tell us jokes the way he did. He was part of us and when he died, all the actions stopped dead and there was no one to do them the way he did.

(Bradbury, 1951)

Abstract

The growing scope and changing nature of academic programs provide a challenge to the integrity of traditional testing and examination protocols. The aim of this thesis is to introduce an alternative to the traditional academic integrity approaches, bridging the anonymity gap, and empowering instructors and academic administrators with new means for maintaining academic integrity that promotes accountability, accessibility and efficiency, preserves privacy, and minimizes disruption to the learning process. This work aims to initiate a paradigm shift in academic integrity practices. Research in the area of learner identity and authorship assurance is important because the award of course credits to unverified entities is detrimental to institutional credibility and public safety. This thesis builds upon the notion of learner identity being comprised of two distinct layers, physical and behavioral, where both criteria of identity and authorship need to be confirmed to maintain a reasonable level of academic integrity. To this end, this thesis is organized in three sections, each addressing one of the perspectives: (a) theoretical, (b) empirical, and (c) pragmatic.

Resum

El creixent abast i la naturalesa canviant dels programes acadèmics constitueix un repte per a la integritat dels protocols tradicionals de proves i exàmens. L'objectiu d'aquesta tesi és introduir una alternativa als enfocaments tradicionals d'integritat acadèmica, cobrint el buit de l'anonimat i donant la possibilitat als instructors i administradors acadèmics de fer servir nous mitjans que permetin mantenir la integritat acadèmica i promoguin la responsabilitat, accessibilitat i eficiència, a més de preservar la privadesa i minimitzin la interrupció al procés d'aprenentatge. Aquest treball té com a objectiu iniciar un canvi de paradigma en les pràctiques d'integritat acadèmica. La investigació en l'àrea de la identitat de l'estudiant i la garantia de l'autoria és important perquè la concessió de crèdits d'estudi a entitats no verificades és perjudicial per a la credibilitat institucional i la seguretat pública. Aquesta tesi es basa en la noció de que la identitat de l'alumne es compon de dues capes diferents, física i de comportament, on tant els criteris d'identitat com els d'autoria han de ser confirmats per mantenir un nivell raonable d'integritat acadèmica. Per a això, aquesta tesi s'organitza en tres seccions, abordant el problemades de les perspectives: (a) teòrica, (b) empírica, i (c) pragmàtica.

Resumen

El creciente alcance y la naturaleza cambiante de los programas académicos constituyen un reto para la integridad de los protocolos tradicionales de pruebas y exámenes. El objetivo de esta tesis es introducir una alternativa a los enfoques tradicionales de integridad académica, para cubrir la brecha del vacío anonimato y para dar la posibilidad a los instructores y administradores académicos a usar nuevos medios para mantener la integridad académica que promueva responsabilidad, accesibilidad y eficiencia, además de preservar la privacidad y minimizar la interrupción al proceso de aprendizaje. Este trabajo tiene como objetivo iniciar un cambio de paradigma en las prácticas de integridad académica. La investigación en el área de la identidad del estudiante y la garantía de la autoría es importante porque la concesión de créditos de curso a entidades no verificadas es perjudicial para la credibilidad institucional y la seguridad pública. Esta tesis se basa en la noción de la identidad del alumno que se compone de dos capas distintas: física y de comportamiento, donde tanto los criterios de identidad como los de autoría deben ser confirmados para mantener un nivel razonable de integridad académica. Para ello, esta tesis se organiza en tres secciones, cada una de las cuales aborda una de las perspectivas: (a) teórica, (b) empírica, y (c) pragmática.

Published Work

Parts of this thesis were previously published by the author as follows:

Amigud, A., Arnedo-Moreno, J., Daradoumis, T., Guerrero-Roldan, A.-E. (2017). Using Learning Analytics for Preserving Academic Integrity. *The International Review of Research in Open and Distributed Learning, 18*(5). doi:10.19173/irrodl.v18i5.3103 (JCR IF=1.732, Q1)

Amigud, A., Daradoumis, T., Arnedo-Moreno, J., Guerrero-Roldan, A.-E. (2018). A Procedural Learning and Institutional Analytics Framework. in S. Caballe & J. Conesa, (Eds.), *Software Data Engineering for Network eLearning Environments: Learning Analytics and Context Awareness Services (Lecture Notes on Data Engineering and Communications Technologies).* Berlin, Germany: Springer-Verlag GmbH.

Amigud, A., Arnedo-Moreno, J., Daradoumis, T., Guerrero-Roldan, A.-E. (2017). A Method for Thematic and Structural Visualization of Academic Content. *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*doi:10.1109/icalt.2017.24

Amigud, A., Arnedo-Moreno, J., Daradoumis, T., Guerrero-Roldan, A.-E. (2017). A Robust and Non-invasive Strategy for Preserving Academic Integrity in an Open and Distance Learning Environment. *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*. doi:10.1109/icalt.2017.23

Amigud, A., Arnedo-Moreno, J., Daradoumis, T., Guerrero-Roldan, A.-E. (2017). Open Proctor: An Academic Integrity Tool for the Open Learning Environment. *Lecture Notes on Data Engineering and Communications Technologies, 262–273*. doi:10.1007/978-3-319-65636-6_23.

Amigud, A., Arnedo-Moreno, J., Daradoumis, T., Guerrero-Roldan, A.-E. (2016). A Behavioral Biometrics Based and Machine Learning Aided Framework for Academic Integrity in E-Assessment. *2016 International Conference on Intelligent Networking and Collaborative Systems (INCoS)*. doi:10.1109/incos.2016.16

Awards

The Doctoral Student Conference Travel Award— sponsored by IEEE Technical Committee on Learning Technology— was presented to the author at the 17th IEEE International Conference on Advanced Learning Technologies for the paper titled "A Method for Thematic and Structural Visualization of Academic Content."

Acknowledgements

Contents

Part II   Empirical Perspective

Part III   Pragmatic Perspective

List of Figures

List of Tables

Acronyms

ACC   Accuracy rate
CRM   Customer relationship management system
DNA   Deoxyribonucleic acid
DT   Decision tree
EDM   Educational data mining
EMS   Education management systems
ERP   Enterprise resource planning systems
ICT   Information and communication technologies
ID   Identity document
IP   The Internet protocol
IT   Information technology
ITU   The International Telecommunication Union
KNN   k-Nearest Neighbor
LA   Learning analytics
LMS   Learning management system
ML   Machine learning
MOOC   Massive open online courses
NB   Naive Bayesian classifier
NLTK   Natural language toolkit
NN   Neural networks
OER   Open educational resources
POS   Parts of speech
PDF   The portable document format
RFC   Request for comments
SIS   Student information system
SQL   Structured query language
SVM   Support vector machine
TF-IDF   Term Frequency-Inverse Document Frequency

Introduction

The expansion of e-learning in higher education has been well noted in the literature (Buzdar et al, 2016). The growing variety of Massive Open Online Course (MOOC) offerings (Salmon et al, 2015) and their ambition to obtain a credit-bearing status (Blackmon, 2016) denotes just that. So does the emergence of the "post-traditional learner," who craves control over how, where, and when to acquire the knowledge (Bichsel, 2013). Maintaining academic integrity becomes an increasingly challenging exercise as physical entities become represented by virtual aliases, when the class size increases, when students are geographically dispersed, and when the teaching and assessment roles become disaggregated. The traditional methods for ensuring the trust relationship stays intact are difficult to translate to learning environments where students and instructors are separated by the time and space gap, and use technology to communicate (Amigud, 2013). These methods stipulate how, when, and where the assessment activities take place and are, at least partly, responsible for the disparity in expectations and experiences of post-traditional learners. When applied to the e-learning context, the traditional strategies negate the very premise of openness and convenience, let alone administrative and economic efficiency. Hence emerges the need for a robust academic integrity strategy

that promotes openness, accessibility, and convenience while allowing for the natural evolution of e-learning toward a more open state.

This thesis explores the issue of academic integrity through the lens of theoretical concepts, quantitative analysis, and practical applications. A behavioral biometrics based and machine learning aided framework is proposed, developed, and applied to create a robust method for aligning learner identities with the work they do in the academic environment. This thesis opens with an excerpt from Ray Bradbury's literary classic that captures a far more efficacious depiction of the notion of behavioral idiosyncrasy than much of the technical definitions do. It also highlights the human ability to perceive and delineate the subtle differences in individual behavior while performing the same actions. The peculiarity of human behavior lies in the core of the proposed method where stylometric and computational techniques are used to capture and differentiate the preferences students exhibit in production of academic content. Another important contribution of this research is the performance baseline of the instructors' ability to classify student-produced content that allows comparisons to be drawn. The effectiveness of academic integrity strategies is underexplored in the literature, hindering the ability to draw quantitative comparisons and ground the decision-making process on data, as opposed to the costs and the sentiment.

Unlike the traditional academic integrity strategies that rely on humans and/or technology to, first, verify learner identities and, second, collect evidence to refute authorship claims, the approach presented in this thesis can concurrently verify both learner identity and authorship claims. Therefore, by minimizing the number of verification steps, the proposed approach aims to provide greater efficiency, convenience, and accessibility.

This research is important and timely, because the changing landscape of learning and teaching presents new challenges, particularly with respect to academic integrity. Unlike instructional approaches, pedagogies, learning technologies, and delivery methods that evolve overtime, the values of academic integrity remain impervious to change and need to be incorporated into the teaching practices, regardless of the mode of instruction or communication technologies used. The credibility and integrity of learning entails an imperative need to establish and maintain a relationship of trust between learners and instructors. It will always be important to know who the students are and to be able to verify authorship of their work. In the absence of these measures, institutions run the risk of issuing course credits to anyone, simply by virtue of participation. Shyles (2002) stressed that "failures to ensure academic integrity and quality control may over time erode institutional credibility, ultimately leading to challenges to accreditation, in addition to a loss of reputation among institutions with high academic standards" (p. 4).

Research objectives

The overarching aim of this thesis is to build upon existing research and develop a robust academic integrity framework that addresses shortcomings of traditional academic integrity approaches. The traditional, observational-based academic integrity strategies are inefficient, expensive, invasive, disruptive, and vulnerable; many of the approaches ignore the behavioral aspects of learning and focus on the cheating behaviors, merely collecting evidence of wrongdoing. A shift in paradigm was needed to address the evolving landscape of education. So emerged the idea to sway away from the traditional physical security paradigm and focus on the behavioral aspects of learning. The learners are the

producers of academic artifacts, and each learner has personal preferences and strategies to piece together the ideas, arguments, and symbols. The challenge was to extract and quantify these behavioral peculiarities and create theoretical and technological frameworks that integrate with the existing processes. The emphasis was no longer on tracking the ways the students cheat, but on the question of how the learners go about producing their work—the way they create academic artifacts. A new framework that simplifies and automates the provision of academic integrity in the e-learning environment—while minimizing administrative overheads, promoting accountability, openness, accessibility, convenience and privacy was needed. To this end, a research program was developed and comprised of the following objectives:

1. Examine literature on academic integrity strategies and identify promising approaches to providing identity and authorship assurance (addressed in Chapter 1)

2. Examine literature on authorship analysis techniques, methods and tools (addressed in Chapter 2)

3. Develop a behavioral-biometrics-based academic integrity framework and test it using real world data (addressed in Chapter 3)

4. Establish a baseline of instructor performance (addressed in Chapter 4)

5. Enhance performance of the proposed framework (addressed in Chapter 4)

6. Integrate research findings into an e-assessment system and develop a prototype application (addressed in Chapter 5)

7. Examine perceptions of instructors toward the proposed academic integrity approach (addressed in Chapter 5)

8. Enhance the functionality of the prototype to serve as an e-assessment platform (addressed in Chapter 6)

9. Design integration framework (addressed in Chapter 7.)

Research questions

This thesis posed the following research questions:

1. What academic integrity strategies can concurrently provide identity and authorship assurance in a convenient and non-invasive fashion? (examined in Chapter 1)

2. What are the techniques for authorship analysis and how well do they perform? (examined in Chapter 2)

3. Could analysis of the student-produced content be used for providing identity and authorship assurance? (examined in Chapter 3)

4. How can classification accuracy be improved? (examined in Chapter 4)

5. How well do instructors perform classification of texts by author? (examined in Chapter 4)

6. What would a system and method look like for maintaining academic integrity based on analysis of learner-produced content? (examined in Chapters 5)

7. What are the instructors' perceptions towards the proposed academic integrity approach? (examined in Chapter 5)

8. What can be done to make the review of written content more efficient? (examined in Chapter 6)

9. What procedural steps need to be undertaken to integrate data-driven approaches into academic and institutional practices (examined in Chapter 7)

Research approach

The research was carried out in five phases. First, a comprehensive review of the literature on identity and authorship assurance strategies was conducted, followed by the review of stylometric and computational techniques. Second, testing was performed to establish viability of the computational approach. Third, a framework was created and developed into a prototype application. Fourth, functionality was expanded to include content-level analysis. Fifth, a procedural framework was developed to provide a roadmap for implementation.

To address a variety of topics and tasks, this research employed a mixed-methods approach and a functional prototyping development methodology. The research commenced with integrative reviews that summarize general themes in existing literature discussing a common issue (Cooper, 1984). It was applied to examine identity and authorship assurance strategies in the academic setting in Chapter 1, authorship analysis techniques in Chapter 2, and analytics frameworks in Chapter 7. The performance of computational and stylometric techniques, as well as the performance of human instructors was quantitatively assessed in Chapters 3 and 4. A survey research methodology was adopted to examine instructors' perceptions towards the proposed academic integrity approach in Chapter 5. A functional prototyping development methodology (Carr and Verner, 1997) was adopted for developing a prototype application presented in Chapter 5 and content visualization module presented in Chapter 6.

Thesis contributions

This thesis makes the following contributions:

1. A comprehensive review of the literature on identity and authorship assurance strategies;

2. A computational academic integrity framework that promotes openness, convenience, and privacy;

3. A baseline of instructor performance for classifying student writings by author;

4. An algorithm for structural and thematic visualization of academic content;

5. A prototype of a behavioral-biometrics based academic integrity system and a modular e-assessment platform;

6. A procedural framework for integrating data-driven approaches into institutional practices.

Thesis structure

This thesis comprises three parts, commencing with the theory, and moving through empirical evaluation to development of the prototype application. The first part of the thesis provides a theoretical perspective and comprises two chapters: Chapter 1 examines identity and authorship assurance strategies, and Chapter 2 discusses techniques and methods for conducting authorship analysis.

Moving from theory to empirical testing, part two of the thesis comprises two chapters. A proof of concept is presented in Chapter 3, followed by examination of the ensemble methods that yield performance improvement in Chapter 4. The notion of idiosyncratic

behavior and more particularly, preferences of language use is central here. In order to establish a comparative baseline, the instructors' ability to perceive differences in the writing styles was quantitatively assessed and compared to the computational methods.

Part three of the thesis takes a pragmatic stance. It commences with a discussion of the development of a prototype system, its features, and architecture; this is found in Chapter 5. Survey results of instructor perceptions towards the proposed approach are examined. The functionality of the proposed system is further enhanced by adding a module for content-level analysis in Chapter 6. This expands the scope of the application from being an academic-integrity tool to a modular e-assessment platform. A procedural-based framework for integrating data-driven approaches is explicated in Chapter 7 using the academic integrity task as an example. The appendices provide a collection of documents supporting the research activities. Appendix A presents the ethics board's approvals, invitations to participate in the studies; and data collection instruments. Appendix B presents a summary of the literature on identity and authorship assurance strategies.

Part I

Theoretical Perspective

This first part of the thesis comprises two chapters. Chapter 1 discusses identity and authorship assurance strategies, their advantages and their disadvantages. It identifies behavioral biometrics as a promising approach to concurrently providing identity and authorship assurance by analyzing patterns in the learner-produced content. Chapter 2 provides a review of the authorship analysis approaches, techniques, and tools that are instrumental to mapping learner identities to their work. The aim here is to introduce the main concepts, approaches, and techniques in both areas and set the stage for the development and evaluation of the computational academic integrity framework presented in part two of the thesis.

Chapter 1

Identity and Authorship Assurance Strategies

Overview

The credibility and integrity of learning entail an imperative need to establish a relationship of trust. It is important to know who the students are and to be able to verify authorship of their work. Throughout the learning cycle, students produce academic content—such as research reports, computer code, portfolios, and forum postings—which serve as the basis for performance evaluation and subsequent credit issuance. Academic institutions have legal and moral obligation to ensure that only students who have completed the work receive the credit. The assurance task is not an easy one to accomplish and has been on academic administrators' radar for over two decades (Amigud, 2013; Crawford and Rudy, 2003; Moore and Kearsley, 1996; Riemenschneider et al, 2016). The challenge stems from the two-tiered nature of academic integrity, comprising identity verification and validation of authorship processes. In other words, one needs confidence in knowing that the students are who they say they are, and that they did the work they claim to have completed. However, confidence comes with a cost; providing both the identity and authorship assurance is a resource-intensive and often intrusive process. As such, academic

integrity is not delivered at a uniform level across all learning activities, but often applied selectively. This approach creates blind spots. For example, assignments submitted electronically may undergo plagiarism screening but do not require identity verification. Similarly, online discussions are generally left unscrutinized, whereas the high-stakes final exams are often proctored, and prior to entering the exam room students are required to present proof of identity.

Strategies that bolster academic integrity may vary in the level of assurance they provide, as well as in their accessibility, cost, ease of implementation, and administration. Strategies that provide greater effectiveness in identifying academic misconduct are often more logistically burdensome to manage, more expensive, and less accessible (Amigud, 2013). Academic integrity strategies can be classified into two classes: those that aim to verify student identities, and those that validate authorship claims. Among the identity assurance strategies are: (a) identity document verification; (b) password-based authentication; (c) biometrics-based identity verification; and (d) challenge question authentication. The list of authorship assurance strategies includes: (a) proctoring; (b) plagiarism detection tools; (c) computer lockdown; (d) network activity monitoring; (d) instructional design; and (e) instructor validation; and (f) policy.

1.1 Identity assurance

The first and primary dimension of learner authentication is identity assurance. It is concerned with verifying that the learners—who are also system users—are who they say they are. Accurate learner identification is critical, because all other components of learner authentication hinge upon on its effective and efficient execution. In the absence

of an effective learner identification strategy, learner progress cannot be tracked. By the same token, the enforcement of academic policies or the detection of academic misconduct is not possible if actions cannot be mapped to physical entities.

Identity assurance strategies may vary from course to course. The quality of identity assurance depends not only on the type of strategy employed, but, more importantly, on the quality of initial identity enrollment. Identity enrollment is an administrative process of acquiring and recording personal information to create an identity profile which needs to be accurately created prior to the identity assurance stage. All subsequent identity verification instances will be compared against the existing identity profile. To ensure quality of identification, an assigned token such as a student number, username or voiceprint needs to be mapped to the legal name and validated against officially issued documents (Bailie and Jortberg, 2008). Identity assurance strategies are summarized in Table 1.1 and explicated in the following sections.

1.1.1 Identity document verification

Traditional and distance schools that conduct proctored exams may verify identity documents before admitting students to take the exam (Amigud, 2013; OReilly and Creagh, 2016; Paullet et al, 2014). This entails administrative overheads as assessment sessions need to be scheduled in advance and managed by human invigilators who are expected to accurately compare students' physical appearance to photographs on their identity documents. Some have argued that administration costs may exceed any potential benefits (Cluskey Jr et al, 2011) and that personal verification using identity documents remains prone to impersonation attacks using fake IDs (Shyles, 2002).

Table 1.1: Summary of identity assurance strategies

Strategy	Advantages	Disadvantages	References
Identity document verification	Common identity verification technique.	Resource intensive not readily scalable requires supervision.	Amigud (2013); Apampa et al (2010); Shyles (2002).
Password-based authentication	Common authentication approach.	Provides low level of identity assurance.	Auernheimer and Tsai (2005); Frank (2010); Ullah et al (2012).
Challenge question-based identity verification	Accessible and convenient. Allows for automation of assessments.	May be disruptive when employed continuously. Requires access to databases of personal information.	Bailie and Jortberg (2009); Barnes and Paris (2013); Jortberg (2010).
Physical biometrics	Provides high level of identity assurance.	Some modalities require special hardware and user interaction; when employed continuously can be computationally intensive and/or perceived as disruptive.	Clarke et al (2013); Rodchua et al (2011); Levy et al (2011).
Behavioral biometrics	Can provide identity assurance in a non-disruptive fashion.	Processing large volumes of data can be computationally intensive. Changes in existing patterns or acquisition of new patterns may require re-enrollment.	Case and Cabalka (2009); Levey and Maynard (2011); Monaco et al (2013).

1.1.2 Password-based authentication

Much of the access to e-learning resources is predominately based on password authentication. Levy et al (2011) stressed that "few schools have implemented students' authentication during online exams beyond passwords" (p.102). Nevertheless, password authentication has its merits. It is highly accessible and does not require any special hardware as is often the case with biometric technology. However, one of its inherent shortcom-

ings is shareability. It allows anyone with the right username and password combination to assume the access rights—and by proxy the identity—of the respective credentials owner. By itself, password-based authentication may not be considered a valid means for maintaining academic integrity or providing identity and authorship assurance. However, it may be used in combination with other techniques, —such as various modalities of biometrics, —that provide a higher level of identity assurance (Auernheimer and Tsai, 2005).

1.1.3 Challenge question-based identity verification

The underlying rationale for using challenge questions is that personally identifiable information should only be known to its rightful owner. Due to its sensitive nature, as it may be used for authentication purposes by financial and governmental services, sharing personal details with unknown parties carries some potential risks. As such, the personal nature of information is expected to serve as a deterrent to sharing. The challenge questions and answers may be user-generated or provided by personal information databases. The former type of challenge questions bears a resemblance to username and password, while the latter lies outside user control and is therefore, subject to the accuracy of information entered thereon.

Some examples of challenge questions include: (a) What is the name of your pet?; (b) What is your mother's maiden name?; (c) What is your date of birth?; and (d) What are the last four digits of your social security number? The first question is user-generated and the answers may be arbitrary, whereas the latter three questions represent personally identifiable data often retrieved and verified against the third-party personal infor-

mation databases. For authentication to be successful, the user must correctly answer the challenge question(s). Nevertheless, this strategy will not be effective in the case of trusted parties or parties such as family members and close friends, who may, already be in possession of the user's personal information. Therefore, identity assurance through challenge questions has its merits and limitations.

Challenge questions may be incorporated into e-learning processes for the identity confirmation of remote learners (Bailie and Jortberg, 2009; Jortberg, 2010; Ramu and Arivoli, 2013; Ullah et al, 2012). A pilot project to test a method of remote learner authentication using personal information was conducted by Bailie and Jortberg (2009). The study also examined students' perceptions of the identity verification process. Learners completed 169 successful identity verifications. Survey results suggest that students may prefer the challenge question verification method in lieu of a proctored environment. Ramu and Arivoli (2013) discussed a secure exam framework that combines keystroke dynamics and challenge questions. Such an approach provides two layers of identity assurance: first, through behavioral biometrics, and the second, through knowledge based authentication. Ullah et al (2012) used a challenge-question based authentication of 39 challenge questions employed continuously to authenticate a group of 13 participants. The results of that study suggest that authentication requests were perceived as inconvenience.

1.1.4 Biometrics-based identity verification

Biometric systems validate physiological and behavioral traits of individuals to confirm their identities. Each biometric trait is unique to the individual and is more difficult to du-

plicate or share than passwords (Levey and Maynard, 2011). To be considered for use in biometric-based authentication, physiological or behavioral attributes need to satisfy the following criteria: (a) universality; (b) distinctiveness; (c) permanence; (d) collectability; (e) performance; (f) acceptability; (g) circumvention (Jain et al, 2004). Biometric technology is often used in commercial access control systems, however their applications are becoming more ubiquitous and starting to emerge embedded in consumer-level products and services. For example, built-in fingerprint scanners can often be found in laptop computers to augment or replace password authentication, and face recognition software in smart phones and photo/video cameras, captures and analyzes facial images for the purposes of user authentication or identification. Biometric systems utilize a wide variety of biometric modalities that include: (a) DNA; (b) ear; (c) face; (d) hand; (e) fingerprint; (f) iris; (g) odor; (h) keystroke; (i) retina; (j) signature; and (k) voice (Jain et al, 2004). The International Telecommunication Union has proposed to delineate physical and behavioral biometrics (ITU, 2012). The former type encompasses physical attributes (who the user is) and the latter type encompasses actions (what the user does). Following this schema, the user's fingerprint geometry (physical attribute) would fall under the physical biometrics category, whereas the user's typing patterns (behavior) would be classified under the behavioral class.

The literature suggests an inclination towards biometrics-based identity assurance strategies. Several studies proposed to adopt physical biometrics, including (a) fingerprint scanners (Auernheimer and Tsai, 2005; Ramim and Levy, 2007); (b) ear recognition software (Rosen and Carr, 2013) ; (c) face recognition software (Irfan et al, 2009; Rodchua et al, 2011), and behavioral modalities that include (a) voice recognition (Mothukuri et al, 2012); (b) keystroke dynamics (Case and Cabalka, 2009; Foster et al, 2009; Monaco

et al, 2013; Ramu and Arivoli, 2013; Stewart et al, 2011); (c) mouse dynamics (McNabb and Maynard, 2010); (d) authorship style recognition (Monaco et al, 2013; Stewart et al, 2011) for the assurance of learner identities.

Behavioral biometric approaches validate behavioral traits that are unique to individuals. McNabb and Maynard (2010) examined an application of mouse dynamics for enrollment and validation of student identities. Mouse dynamics is a behavioral biometrics approach that aims to verify user identity on the basis of mouse movements. Seventy-three students from the University of Texas System TeleCampus participated in the study and completed a survey. The results suggest that user authentication through mouse dynamics was perceived by much of the students as easy to use. In a similar study, Case and Cabalka (2009) discussed a secure testing solution that employs video monitoring and keystroke dynamics. Similar to mouse dynamics, keystroke dynamics aims to verify user identity on the basis of typing patterns.

There are several tools for enhancing security of e-assessments whose features include biometric authentication (OReilly and Creagh, 2016). A non-exhaustive list of these tools is depicted in Table 1.2. For an analysis of biometric technologies, advantages and disadvantages of biometric-based authentication, please refer to (Unar et al, 2014). In spite of biometrics being proposed as a solution to learner identification in much of the literature, the number of commercial applications that employ biometric technologies for learner authentication is limited. So is the variety of modalities employed, which include: (a) fingerprint geometry; (b) facial geometry; (c) voiceprint; (d) mouse dynamics; (e) keystroke dynamics; and (f) stylometry. Unlike the password-based authentication that will only accept the correct username and password combination, biometric technologies are prone to generating both false positives and false negatives. Knowing how well these

technologies perform and how to interpret the reports produced by these tools is important to avoid making procedural errors. This applies to automatic proctoring products as well. Colwell and Jenks (2005) stress that "[T]he accusation or allegation that students cheat must be made with extreme caution. Many students would be severely offended if they were accused of any type of impropriety" (p.17). However, performance data is not always readily available on the vendors' websites, which suggests a need for further investigation.

There are benefits and challenges that come along with biometric authentication. On the one hand, it provides greater identity assurance; on the other hand, its efficiency depends on the stability of the biometric features. For example, a sports injury may significantly affect an individual's typing abilities and therefore the ability to complete authentication through keystroke or mouse dynamics. It could be reasoned that such authentication measures require redundancy to ensure accessibility. Some scholars propose combining multiple biometric traits to maintain continuity and increase process effectiveness (Apampa et al, 2010; Kang and Kim, 2015; Rabuzin et al, 2006).

To maintain the security of e-assessment activities, it is equally important to know who the learners are as it is to know whether their authorship claims hold truth. Because biometric authentication establishes who the learners are, in the context of e-learning its applications most often need to be augmented by another technique such as video monitoring to establish what learners do. Authorship assurance strategies will be discussed in the following sections.

Table 1.2: Biometric-aided tools

Vendor	Website	Strategy	Biometrics	Identity Assurance	Authorship Assurance
SoftwareSecure	softwaresecure.com	Remote Proctoring	Fingerprint	Physical biometrics; photograph	Monitoring; usage restrictions
Kryterion Inc.	kryteriononline.com	Remote Proctoring	Keystroke dynamics, Facial recognition	ID check over webcam; Multimodal biometrics	Monitoring; usage restrictions
BSI	biosig-id.com	Biometric authentication	Gestural biometric	Behavioral biometrics	–
B Virtual	bvirtualinc.com	Remote Proctoring	Gestural biometric	ID check over webcam; biometrics	Monitoring; usage restrictions; Screen sharing
Voice Proctor	voiceproctor.com	Remote proctoring	Voiceprint	ID check over webcam; Random voice verification over the phone	–

1.2 Authorship assurance

The second and equally critical factor in the learner authentication process is authorship assurance. Learners demonstrate mastery of subject matter through a range of learning activities involving both formative and summative assessment. The process of authorship assurance is concerned with establishing the veracity of authorship of academic content produced by a learner as a result of a learning experience. At the core of this process is the question of what artifacts have learners produced as a result of the learning experience? "As online courses are growing in popularity, more and more instructors are skeptical of whether or not the work submitted is actually completed by the student who is enrolled" (Hoshiar et al, 2014, p.338). And these concerns are not without merit. The media attention surrounding cases of academic misconduct in institutions of different sizes and types denotes just that. Survey data from 42 Canadian universities suggests that in the 2011-2012 academic year, over 7,000 students were facing disciplinary action for academic

misconduct (Moore, 2014). Authorship assurance follows the identity assurance process and maps authorship claims to the learner's identity. Table 1.3 presents a summary of the authorship assurance strategies which are explicated the following sections.

The detection of cheating is not always a clear-cut exercise. When a source paper is presented to prove a case of plagiarism, it removes the least shadow of doubt and the evidence is difficult to refute (Colwell and Jenks, 2005). However, the detection of collusion, custom written papers, and exam sharing appears to be a more challenging task. Authorship assurance strategies can be divided into two types: (a) evidence based and (b) performance based. Much of the authorship assurance strategies are of the former type and concerned with gathering evidence to refute student's authorship claims. When no evidence is found, student's work is deemed original. The latter type is concerned with verifying consistency of content production and requires a baseline. Authorship enrollment is a process of establishing a baseline of the learner's current academic abilities and preferences. When learners apply to their programs of study, they may be required to complete an entrance exam or provide a sample of their prior academic work. Throughout the course of study, learners produce content. This information becomes part of the student profile. Enrollment is a continuous process. Student acquire new knowledge and competencies and experience academic growth; their vocabulary and writing style may evolve overtime. These changes need to be reflected and the baseline reestablished.

1.2.1 Plagiarism detection tools

Plagiarism detection tools analyze written materials such as text documents and computer source code for duplicate content found in its database, or external sources. They

Table 1.3: Summary of authorship assurance strategies

Strategy	Advantages	Disadvantages	References
Plagiarism detection	Automated reporting. Accessible and convenient.	Resource intensive; not readily scalable requires supervision.	Culwin (2008); Fiedler and Kaner (2010); Kermek and Novak (2016).
Traditional proctoring	Suitable for any assessment task.	Resource intensive; requires physical attendance employs human invigilators. May affect convenience and/or accessibility due to scheduling and travel requirements.	Apampa et al (2010); Cluskey Jr et al (2011); Shyles (2002).
Remote proctoring	Convenient; eliminates travel requirements. Suitable for any assessment task. Can be automated.	Resource intensive; requires human intervention to monitor live or review recorded sessions. Affects privacy and may require grant of access to third parties for monitoring learner's personal computer and/or assessment environment.	Case and Cabalka (2009); Foster et al (2009); OReilly and Creagh (2016).
Behavioral biometrics (learner-produced analysis)	Provides identity and authorship assurance in a single tiered process. Non-intrusive and convenient. Allows for automation of assessment tasks.	The scope of assessments is generally limited to written content. Can be computationally intensive for large datasets. Changes in existing patterns or acquisition of new patterns may require reenrollment.	Monaco et al (2013); Stewart et al (2011).
Instructor validation	Supports continuous assessment. Promotes the values of trust and accountability.	Not scalable to massive course level; time consuming task.	Barnes and Paris (2013); Saunders et al (2008); Shyles (2002).
Computer lockdown	Supports various types of assessments. Can be combined with other techniques.	Affect convenience and /or privacy; May require installation of software when employed on learner's personal computer.	Chiranji et al (2011); Lilley et al (2016); Rodchua et al (2011).
Network monitoring	Can be combined with other techniques. Suitable for any e-assessment task.	Not designed to validate authorship but to dispute authorship claims.	Chiranji et al (2011); Gao (2012); Pan et al (2004).
Instructional design	Accessible and convenient; built into the course design.	Aims to minimize cheating; Provides low level of authorship assurance.	Chiesl (2007); Fendler and Godbey (2015); Mott (2010).
Policy	Explicitly states guidelines; easy to disseminate; is meant to be educational tool.	Requires infrastructure for issue tracking and enforcement.	Amigud (2013); Gullifer and Tyson (2014); Murphy and Holme (2015).

are not effective in attributing authorship, but to the contrary, they aim to find evidence to negate student's authorship claim. There are both free and commercial plagiarism checkers available. Instructors may also use popular search engines as a first line of defense against plagiarism (Amigud, 2013) or use software such as JPlag, Plaggie, Moss, Sherlock (Kermek and Novak, 2016) to conduct the analysis. However, plagiarism detection tools have limitations and may not detect all non-original content in learners' papers (Culwin, 2008; Fiedler and Kaner, 2010).

Gregory and Strukov (2002) conducted an experiment consisting of 8 conditions to assess the limitations of the plagiarism detection tool Turnitin. The results suggest that the service exhibited limitations in its ability to detect and flag plagiarized content from the following source materials: password-protected sites, PDF files, and foreign languages, with the exception of some Spanish-language content. The findings tally with that of Fiedler and Kaner (2010), who compared the effectiveness and perceived effectiveness of two services Turnitin and MyDropBox, by conducting a survey of 954 deans. The results suggest that plagiarism detection tools have blind spots. These services may not have access to all restricted databases of scholarly articles, and therefore, are not always able to index their content. Various detection tools provide different plagiarism scores for the same content (Fiedler and Kaner, 2010). The findings are in line with that of Culwin (2008) who stressed that "[u]sing the tool on the same document at different times may give dramatically different outcomes" (p.190). In another study, Heckler et al (2013) examined whether the plagiarism detection tool Turnitin, can serve as a deterrent to cheating behavior. The results suggest that students who were made aware of plagiarism screening exhibited lower incidence of plagiarism. Kermek and Novak (2016) proposed a framework for detecting plagiarism of source code used in programming assignments and con-

ducted a study to assess its performance. The results suggest a promising direction and further research is required to assess its performance on larger and more diverse data sets.

1.2.2 Proctoring

Much of the testing and examination in traditional courses are conducted in a proctored setting. Proctors verify identity documents, keep student attendance, and keep a close watch on examinees. Identity fraud and cheating is still possible if identity documents are falsified or students collude with proctors or peers, although risks are arguably lower than that of non-proctored environment. Teaching assistants and professors may be present at the exams, thus adding another level of assurance against proxy test-takers. OReilly and Creagh (2016) discussed variations in the approaches to proctoring that can be organized into three categories that include: (a) traditional proctoring; (b) technology-facilitated proctoring; and (c) automatic proctoring. The first two types rely on human invigilators to detect misconduct, while the last is applying pattern recognition techniques to detect anomalies in an automatic fashion.

Remote proctoring is often proposed as a means to control the remote learning environment (Apampa et al, 2010; Chiranji et al, 2011; Frank, 2010; Mothukuri et al, 2012; O'Reilly and Creagh, 2015; OReilly and Creagh, 2016; Rodchua et al, 2011; Rosen and Carr, 2013). Three pilot projects of remote proctoring technology have been conducted at Pennsylvania State University World Campus by (Foster et al, 2009), at Western Governors University by (Case and Cabalka, 2009), and at the University of Hertfordshire by Lilley et al (2016). The results suggest that remote invigilation may constitute a viable alternative to traditional proctoring.

However, the need of remote monitoring may not be universally shared. A study by Saunders et al (2008) assessed the perceptions of 56 chairpersons of accounting departments at universities and colleges in the U.S. on the issues of academic integrity and identity assurance. The majority of participants (35%) did not agree with the necessity to employ webcam monitoring for course registration, whereas nearly 25% of the participants agreed with it. The findings tally with that of Cluskey Jr et al (2011), who share a skeptical view of proctoring; they noted "we believe that costly proctor supervision provides only minimal assurance of academic integrity" (p.3). Privacy is another area of concern, as some learners may perceive risks in allowing remote monitoring of their learning environment and granting access to their personal computers (Lilley et al, 2016).

1.2.3 Behavioral biometrics

Students interact with ICT and learning content (Moore and Kearsly, 2005) which results in abundance of mineable data. Students exhibit different preferences of language use and the use of input devices, such as the keyboard. The rationale for using behavioral biometrics for authorship assurance is that actions performed by the same student are expected to show greater similarity than that of different students.

Stylometry is the term used to describe the process of measuring quantifiable stylistic features for the purposes of authorship analysis. Monaco et al (2013) define stylometry as "the study of determining authorship from the authors' linguistic styles. Traditionally, it has been used to attribute authorship to anonymous or disputed literary documents" (p.1). Stylistic preferences exhibited by authors, are thought to be author-specific, where predictions can be made based on the presence or absence of certain linguistic elements

in a text. When applied to the e-learning context, this approach is unique in that learner identity is derived from authorship which makes it a one-tiered application to attain both identity and authorship assurance. For more information on trends in authorship analysis, please refer to PAN/CLEF evaluation lab (Stamatatos et al, 2015). An experimental study conducted by Monaco et al. (2013), building upon the work of Stewart et al. (2011), compared performance of keystroke dynamics and stylometric analysis. The results are promising, although keystroke dynamics slightly outperformed the stylometric approach. Both the keystroke dynamics and stylometry behavioral biometrics have several benefits. They can concurrently provide identity and authorship assurance and are accessible, because they work with the standard computer keyboard. Moreover, data collection and processing can be automated and performed on the background without a need for user interaction. However, one key difference exists between the two techniques is that stylometry does not require continuous verification of user input, therefore, content can be produced off-line using tools and methods that students feel comfortable using.

1.2.4 Instructor validation

Instructors play an important role in the validation of student work. Colwell and Jenks (2005) stressed that "[I]t is the duty of every online instructor to be mindful of the requirement to make the course a credible evaluation of the students' knowledge" (p.2). Instructors can be considered human biometric scanners capable of evaluating the authorship and identities of the learners they teach. Barnes and Paris (2013) maintained that "[e]xperienced instructors know one key to recognizing cheating or plagiarism is to become familiar with a student's writing style" (p.4). In a study by Fiedler and Kaner (2010)

authors suggest checking specifically for plagiarism "by adopting a skeptical mindset toward the papers before taking a second pass through remaining papers in which you treat them as honest submissions deserving of feedback" (p.43). This is especially relevant to graduate programs, where annual enrollment is limited to a small number of students. Students work closely with professors, enabling the latter to track academic progress and examine both quality and authorship of the submitted materials. However, this approach may not prove effective in learning environments where instructors do not get a chance to know their students, particularly where the class size is measured in hundreds.

To keep up with the advancement of technology and growing class sizes, instructors may require additional tools to help them identify learners and validate the originality of their work. Some studies suggest that instructors may be reluctant to address the issue of academic misconduct (Fendler and Godbey, 2015; Heckler et al, 2013). The effectiveness of instructor validation is yet to be determined.

1.2.5 Computer lockdown and network monitoring

Authorship assurance through monitoring and control of the remote learning environment is a strategy comprised of two different techniques: (a) deterrence of cheating behavior by imposing access restrictions to external resources; and (b) activity monitoring to identify such risks as collusion and proxy test-taking. The rationale behind this approach is that by minimizing the opportunity for access to the external resources during testing and examination, the opportunity to cheat is reduced (Rodchua et al, 2011). Similar to other strategies, the integrity of the assessment is presumed to be valid unless there is evidence to the contrary. Chiranji et al (2011) proposed a comprehensive approach to secure online

examination through audio-visual monitoring of the examination environment combined with access restriction to external information sources and network traffic monitoring. A study by Pan et al (2004) described a framework for conducting secure electronic assessments within the existing academic network environment through the application of a distributed firewall that monitors user activity during the exam session. A paper by Gao (2012) reviewed common academic integrity strategies and proposed a method of collusion detection using IP monitoring. They argue that the probability of two students from the same location taking the same courses is small; therefore, any coincidental similarities in IP addresses and time are indicative of possible academic misconduct. Location verification is a technique used in remote proctoring applications (OReilly and Creagh, 2016).

1.2.6 Instructional design

Instructional design is a set of techniques and approaches for the design, development and delivery of learning materials and activities. The rationale behind this strategy is to increase the costs and decrease the benefits of cheating behavior by assigning conditions to learning activities such as setting exam time limits, randomizing test questions, and allowing multiple exam retakes (Chiesl, 2007). However, unlike the proctoring or monitoring strategies that attempt to discover instances of cheating, authorship assurance through instructional design attempts to manipulate learners to refrain from academic misconduct by virtue of the assessment format. Some scholars suggested using a large pool of question sets (Rodchua et al, 2011) and conducting a large number of exams during the semester with a low grade weight (Chiesl, 2007). Some have argued that a student

may seek the help of a proxy test-taker and cheat on one assessment; however, if there are 10 assessment activities in a semester, the sheer volume should discourage the cheating practice (Chiesl, 2007). Others highlighted the benefits of test question randomization (Chiesl, 2007; Mott, 2010; Ramu and Arivoli, 2013), which takes away the benefit of predictability. Some scholars have argued for the imposition of time limits during assessment activities (Barnes and Paris, 2013; Chiesl, 2007; Cluskey Jr et al, 2011). "A tight time frame will discourage students from cheating. Students will barely be able to complete the exam and will not have time to thumb through the text looking for answers" (Chiesl, 2007, p. 206).

1.2.7 Policy

Honor codes and academic integrity policies serve as both an educational tool and a deterrent. They emphasize the importance of the values of trust and integrity or outline penalties for violation of trust. However, some studies question their effectiveness. In one example, the results of a survey of 3,405 students conducted by Gullifer and Tyson (2014) suggest that only half (52%) of surveyed students had read the academic integrity policy. Interestingly, students in face-to-face mode were less likely to read the academic integrity policy than students enrolled in a distance course. A study by Beasley (2014) examined learners' perspective on prevention of academic misconduct and noted that ignorance of consequences and rules was cited as one of the reasons students engage in cheating. Some have argued that the inconsistent use of the term "plagiarism" creates confusion among both instructors and students leading to inadvertent acts of plagiarism (Gullifer and Tyson, 2014). A continuous dialogue between learners and instructors on

the issues of trust and integrity may clarify any ambiguities the learners have, however communication by itself is not sufficient to achieve compliance. What follows is the need for additional layers of assurance against academic misconduct. But even when secure assessment techniques are available to the faculty and administrators, they need to be consistently applied in order to mitigate the risks (Paullet et al, 2014).

1.3 Data analytics-aided methods

Throughout the course of assessment activities, learners interact with content, computer systems, and environment (Moore and Kearsly, 2005). These interactions are often recorded in a form of access logs, video and audio content (if remote proctoring is employed), and academic artifacts that are the main staple of the assessment exercise. The analysis of these interactions may help to delineate activities that contravene academic integrity standards. This is the underlying principle behind strategies that employ analytics to bolster e-assessment security. Some of the approaches that employ data analysis techniques to maintain academic integrity include: (a) non-oiginal content detection (Culwin, 2008; Fiedler and Kaner, 2010; Kermek and Novak, 2016); (b) network activity monitoring (Chiranji et al, 2011; Gao, 2012); (c) keystroke dynamics identity verification (Case and Cabalka, 2009); (d) mouse dynamics identity verification (McNabb and Maynard, 2010); (e) stylometric identity verification (Monaco et al, 2013; Stewart et al, 2011); (f) authentication through physical biometrics (Clarke et al, 2013); and (g) automated proctoring (O'Reilly and Creagh, 2015).

Data analysis may be conducted using statistical methods and/or machine learning techniques. While the former is more familiar to the researchers using quantitative meth-

ods, the latter may benefit from a short introduction. Machine learning is a branch of computer science that encompasses computational theories and methods for discovering patterns in data. Its advantage over statistical modeling is that machine learning algorithms learn from data, eliminating programming requirements. Analyses can take the form of supervised or unsupervised learning. In an unsupervised learning task, a single set of data is used, which is then passed on to an algorithm to discover patterns within. In a supervised machine learning task, data (a training set) get passed on to an algorithm that builds a model. The model is then applied to a new set of data (a testing set) which organizes the unseen data according to features in the training set. Much of the e-assessment security issues can be tackled using either statistical modeling or machine learning techniques depending on the type of data and task. For further information about machine learning, please refer to (Domingos, 2012).

One key advantage of using analytics for the provision of security of assessment activities is the ability to automate tasks. Automated proctoring is a good example of a task that requires multiple data analyses to be performed concurrently, making it computationally and resource intensive strategy. For instance, an audio-visual stream from a learner's webcam may be analyzed to authenticate the learner's face, track the learner's eye movements, identify changes in lighting conditions, or identify the presence of other people. Similarly, input devices may be monitored to detect shortcut presses and changes in typing patterns (O'Reilly and Creagh, 2015), and textual data can be analyzed to find author-specific patterns indicative of the author's identity (Monaco et al, 2013; Stewart et al, 2011).

Non-original detection tools used for the detection of plagiarism in text and source code also use algorithms that calculate similarity/distance between content items. At least

two studies have examined the issue and proposed frameworks for addressing the issue of plagiarism in programming assignments (Bejarano et al, 2013; Kermek and Novak, 2016). Both studies have shown promising results and involved experimental work and prototype development. These approaches automate assignment screening for plagiarism however, manual intervention would still be required to control for type I and II errors. The main limitation of these approaches is that they are only directed at disproving authorship claims.

1.4 Implications for learner assessment

The advantages of e-assessments over traditional pen and paper exams have been noted, and some maintain that traditional exams are being gradually replaced by electronic assessments (Ramu and Arivoli, 2013). A decision to select one form of assessment over another is often driven by institutional policy. Gibbs and Simpson (2004) noted that:

> Resource constraints in conventional universities have led to a reduction in the frequency of assignments, in the quantity and quality of feedback and in the timeliness of this feedback. Modularization has tended to shorten courses and has reduced the timescale within which it is possible to set assignments and provide feedback, while increasing the number of examinations. (p.9)

Much of the institutions with distance programs had adopted a contrary policy, decreasing a number of high-stakes examinations and substituting them with alternative forms of assessment such as assignments, projects, and portfolios (Bailie and Jortberg, 2008) where secure examinations were not feasible. These alternative assessments are often conducted in a continuous fashion. Continuous assessment entails more frequent interac-

tion between the student and instructor, where evaluation and feedback follows each unit of learning. Therefore, arguably, this approach has advantages in allowing instructors to familiarize themselves with learners' academic skills over the on-demand assessment approach that lacks frequency of interaction and feedback. It would be safe to assume that unlike on-demand assessment, its continuous counterpart is more resource intensive due to the higher volume of assessment activities. ICT could provide a path to more efficient continuous learner-instructor communication; however its success is contingent upon the organization's technical expertise and financial resources. Some of the implications of identity and authorship assurance on continuous assessment include (a) a change in the level of interaction between learners and instructors and; (b) the role of instructor; (c) a need for the automation of assessment activities; and (d) higher demand for support services. Lopez et al (2007) maintained that implementation of continuous assessment requires additional efforts on the parts of faculty and students. It is likely to require course content to be reviewed daily, despite the temptation to study closer to exams. The volume of formative assessment may also increase; however, it may be integrated into continuous assessment tasks. As such, learning and teaching styles would need to be adjusted. Instructor validation has its merits and may provide the first line of defense against academic misconduct; however, this approach is not readily scalable on a massive course level. Whether assessment is on-demand or continuous, instructors are on the frontline, dealing with academic misconduct cases. Some research suggests a lack of commitment on the part of the faculty to enforce academic integrity either due to time and resource constraints (Heckler et al, 2013), difficulty presenting compelling proof (Fendler and Godbey, 2015), or the legal and privacy issues (Murphy and Holme, 2015).

It can be contended that when instructors and academic administrators are empowered with technology that aligns learner identities with their academic work and provides a reasonable level of assurance and automaticity while imposing minimal disruption on the learning process, the process of managing academic misconduct cases becomes more efficient. Some have argued that better monitoring could contribute to the reduction of cheating behavior (Heckler et al, 2013). Strategies based on behavioral biometrics may be found to be particularly useful. These technologies are suitable for aligning learner identities with academic work they do in a one-tiered approach and in a non-intrusive fashion. Behavioral biometric techniques have advantages over their physical counterparts in that they do not require specialized hardware and therefore offer greater accessibility— an important factor when considering technology selection. The quality of authentication is determined by the system's ability to accurately match the presented biometric sample to the existing identity data. Such an approach has limitations and is prone to the type I and II errors; thus, multiple authentication factors may be employed to ensure a fail-safe operation and increase the level of assurance. Furthermore, in order to maximize control over the examination environment, multiple identity and authorship assurance strategies may be combined together to account for as many environmental variables as possible. However, multi-tiered strategies vary in the level of user interaction and the level of security they deliver. Many of the current learner authentication strategies are intrusive and therefore may provide cues to those engaging in cheating behavior as to when the required verification will take place (Clarke et al, 2013). Non-intrusive learner authentication techniques such as face recognition, keystroke dynamics, and strylometry behavioral biometrics are not as disruptive, but may have implications for user privacy.

1.5 Privacy implications

Formal learning entails a contractual agreement between learners and institutions that requires adherence to academic policies and administrative procedures—where academic integrity principles comprise a subset. Collection, access to and processing of learners' personally identifiable information is necessary and justified on the grounds that course credits cannot be issued to anonymous entities and similarly, academic policies cannot be enforced in an anonymous environment. Therefore, formal learning cannot function behind a veil of anonymity. As such, validation of identity and authorship information will remain an imperative part of the academic process. Privacy and data security will become increasingly important issues as academic integrity strategies evolve to integrate technologies that utilize new forms of personally identifiable information and employ data collection methods that do not require user interaction. The challenge will be to strike a balance between security, privacy, and convenience. The level of adoption and ubiquity of these strategies will depend on how this information is managed by service providers and how its level of security is perceived by their owners. Aceves and Aceves (2009) stresses that "[t]he overarching concern of student authentication and identity management is much larger than distance education; it is an issue that our institutions are struggling with as state and federal laws protecting data privacy in a technology-driven environment become stricter" (p.147). As new security technologies enter the academic arena and collect new types of personal information, legislation might not be always rapid enough to keep up with the pace of technological progress, which could result in an iterative cycle of discovering new threats to privacy and attempts to minimize them. This has an impact on the way academic integrity measures are delivered. Presently, for

example, evidence of cheating may not be collected directly from the learners' mobile devices due to the privacy laws (Murphy and Holme, 2015); however, mobile devices may be banned from exams, and having one in possession during an assessment session may be considered sufficient evidence for disqualification regardless of its use. Such a policy will be difficult to enforce in the distance learning environment. Although smart phones and other mobile technology can pose a threat to exam integrity, these devices can be used as tokens (what the user has) for device-based authentication or to establish a proof of presence.

Jiao (2011) has argued that the scope of learners' privacy is not limited to contact information and should also encompass information pertaining to learners' academic skills. This is an important point considering that behavioral biometric technologies collect data relating to the level of learner performance. It could be reasoned that the use of learner-generated content beyond the scope of secure assessments should raise a privacy concern; however, learning is often an open process and much of the academic artifacts such as assignments and portfolios are published with the very purpose of publicly demonstrating the level of competence or skill. Learners' computers often contain personal data, as computers are used for activities other than learning. The use of proctoring services, however, may require the user to install a proctoring application that allows human invigilators to remotely disable certain features, share the screen, and log keystrokes, among other functions. Lilley et al (2016) conducted a study to examine learner perceptions towards the use of remote proctoring. They noted that prior to the participation in the study some participants expressed concerns about data protection and privacy. The concerns stemmed from the need to share personal information, allowing access to the personal computer, and enabling a live video feed to a "stranger" (p. 3). The use of physical bio-

metrics may also give cause for concern particularly if the data is managed by a third party. Levy et al (2011) conducted a survey of 163 online students enrolled on an introductory IT course to examine their perceptions of sharing biometric data. The findings suggest that online learners are more likely to enroll their biometric profile and use biometric authentication through their university than they are to enroll the same credentials through a third-party service provider.

1.6 Summary

This chapter examined a variety of identity and authorship assurance strategies for enhancing academic integrity, their advantages, and the shortcomings, as well as implications for assessment and privacy. As was discussed in the previous sections, some learner authentication strategies are well suited for identity assurance only, some are designed to validate authorship claims, while others validate both identity and authorship. Some approaches rely on technology, others emphasize instructional design, and some require invigilation by humans. In the absence of sufficient measures for providing a reasonable level of authorship and identity assurance, whether in the traditional school environment or online, academic integrity could be undermined and institutional credibility could be impacted. After all, the identity of a learner and authorship of a paper submitted in-person to an instructor who only facilitates the lecture is no different from one submitted electronically, unless, of course, the instructor employs additional measures to validate the learner's identity and authorship. Perhaps the brick-and-mortar schools may benefit from the expertise of distance institutions and join efforts toward the development of a robust model of identity and authorship assurance. Distance institutions have been exploring

the alternatives to traditional proctoring and on-demand assessments, and that allowed them to take the lead in the research of academic integrity approaches. A part of the solution may lie in the perpetual reinforcement of academia's core values of trust and integrity (Culwin, 2008; Lang, 2013). Continuous validation of identity and authorship claims provides just that. One study suggests that learners who were reported for academic misconduct pushed blame on instructors for not proctoring their work (Beasley, 2014). When instructors are armed with technology that aligns learner identities with their academic work, opportunities for excuses are no longer provided. If the cost-benefit balance of cheating is disrupted by lifting the pressure of on-demand assessments from the students' shoulders and replacing it with a continuous approach that promotes a more gradual mastery of a subject matter and allows for second chances, the need for a controlled testing environment may be less pressing.

Technology plays an important role in providing an efficient means of identity and authorship assurance when there is a physical gap between learners and instructors. From the list of available options, behavioral biometrics technology— and stylometry in particular— presents a promising approach to providing academic integrity in the e-learning environment. The key advantages of stylometry are that it can concurrently perform verification of identity and authorship by analysing patterns in the student-produced content, it does not require special hardware, and it is capable of collecting data with minimal disruption to the learning and teaching processes. The next chapter provides a review of the stylometric techniques, methods, and tools that will be used as input for the development of the academic integrity framework in Chapter 3.

Chapter 2
Authorship Analysis

Overview

The previous chapter discussed approaches to academic integrity, where behavioral biometrics—applied to the analysis of the student produced content— emerged as a promising approach to providing concurrent identity and authorship assurance. The advantage of stylometry over keystroke or mouse dynamics is that data collection does not need to be continuous. It takes advantage of the abundance of student-generated content to verify student identities and validate their authorship claims in a one-tiered process, without the need for special hardware. This chapter explores this notion further and provides an overview of methods and tools for analysis of patterns in the textual data.

The notion of authorial discrimination has its roots steeped in history. Some of the early examples of textual classification date back to the Hellenistic period where the works of influential Greek authors were organized by genre as well as compared on the basis of their structure (Love, 2002). Through this time period there are documented examples of authorial attribution and plagiarism disputes. In one example, Aristophanes of Byzantium, the librarian at Alexandria, was able to identify plagiarism during a poetry competition and corroborate his argument by presenting the original texts as evidence

(Love, 2002; McGill, 2012). The notion of style as a distinguishing attribute of authorship, and further quantification of stylistic features, emerged in the 19th century. Holmes (1998) maintained that early attempts to quantify the writing style may be traced back to Augustus de Morgan, who in 1851, suggested using word length as a stylistic marker. The question of who wrote what has been puzzling scholars in their search for historical facts and prompted a number of authorship analysis studies. Holmes (1998) noted that Mendenhall (1887) used word-lengths as a discriminator of authorship to examine the works of Shakespeare, Yule (1938) used sentence length to examine the authorship of De Imitatione Christi, while Cox and Brandwood (1959) chose to discriminate authorship on the basis of the distribution of the last five syllables of each sentence in the works of Plato.

With the progression of time, the nature of inquiry shifted from resolution of literary disputes to solving pragmatic issues (Stamatatos, 2009). Today, authorship analysis techniques aid a forensic investigation (Johnson and Wright, 2014); they are used to identify cases of plagiarism (Alsallal et al, 2013) and internet-facilitated social misconduct (Stein et al, 2009); and even facilitate continuous user authentication (Canales et al, 2011). The large volume of textual data entails a need for an efficient data processing approach which fueled research interest in information retrieval, natural language processing (NLP), and machine learning (ML) techniques that provide the necessary tools to carry out textual analysis (Stamatatos, 2009). The former deals with storage, retrieval, and classification of large volumes of textual data. The latter addresses the issue of quantification and translation of the meaning of human readable content into machine code for processing. The last enables computers to carry out prediction, clustering, and classification tasks without being explicitly programmed. The advancements in data processing created a favorable

ground for the development of computational techniques that precipitated a paradigm shift, advancing stylometric research from statistical similarity analysis to pattern recognition tasks.

Although, many of the applications for authorship analysis bear a forensic character, they should not be viewed as strictly security-oriented. For example, user authentication and academic plagiarism detection are some of the uses to which authorship analysis techniques can be put outside of the legal sphere. The aim of this chapter is to introduce key concepts and stimulate thinking about stylometry as an interdisciplinary field of study. Much of the analyses have been conducted using different methodological approaches from within individual disciplines, such as statistical stylistics, forensic linguistics, and computer science, just to name a few.

2.1 Definitions

The term authorship analysis serves as an umbrella term that comprises authorship identification, authorship verification, and author profiling (Brocardo et al, 2013). Authorship attribution, also termed recognition or identification deals with matching an "authorless" text to a corpus of written works of known origin. It has a one-to-many relationship and aims to identify the most likely author of a written sample from the list of available authors. Authorship verification or authentication is a binary process which compares two text samples. It has a one-to-one relationship and attempts to prove or disprove the likelihood of two texts being written by the same author. Author profiling is concerned with correlation and prediction of personal characteristics such as gender, age, personality type, etc.

Quantitative measurement of stylistic features—often referred to as style markers or discriminators — lies in the core of authorial analysis. Style markers are linguistic elements, the occurrence of which varies among authors, is thought to be habitual and outside of conscious control (Brennan et al, 2012; Brocardo et al, 2013), and considered a form of behavioral biometrics (Fridman et al, 2015; Monaco et al, 2013; Saevanee, 2015) sometimes compared to a fingerprint (Iqbal et al, 2013). The presence of certain style markers may also serve as the basis for classification by locale, age and gender (Farias et al, 2013) as well as the personality type (Noecker et al, 2013).

Stylometry—the measurement of style—is the term often utilized in the literature to refer to the mechanism behind feature selection, quantification and measurement. One definition of stylometry is "the use of numerical methods for the solution of literary problems, most often problems of authorship, integrity, and chronology" (Michaelson and Morton, 1972, 89). Another definition is "statistical analysis of literary style [that] complements traditional literary scholarship since it offers a means of capturing the often elusive character of an author's style by quantifying some of its features" (Holmes, 1997, 1). A definition that carries a modern connotation, is "a behavioral feature that a person exhibits during writing and can be extracted and used potentially to check the identity of the author of online documents" (Brocardo et al, 2013, 1). This thesis defines stylometry as the process of measuring quantifiable stylistic features for the purposes of authorship analysis. All of these definitions share a common denominator, namely the linguistic elements which are peculiar to individual writers and computational methods devised to measure them.

2.2 The evolution of stylometric methodology

Over the past 40 years, stylometric methodology has undergone a change, which is apparent from the definitions. The 1972 definition of stylometry by Michaelson and Morton (1972) emphasizes literary disputes, whereas the 2013 definition by Brocardo et al (2013) places emphasis on identification of electronic texts. As technology moves forward, the methodology incorporates more advanced approaches. This section provides a brief overview of the evolution of stylometric methodology which has undergone a transformation from being a manual effort conducted by experts examining paper-based documents, to a computer-mediated process that offers some degree of automation.

Traditional authorship analysis methods rely on the knowledge of human experts who examine the context and linguistic factors influencing variation of authorial styles. It entails a holistic view of the authorship process and combines the attributes of qualitative research. However, its successor, the computational-stylometric paradigm (Chaski, 2005) encompasses the current non-traditional authorship analysis techniques and limits the scope to analysis of stylistic markers. The quantitative approach takes a reductionist stance, leaving out the context, such as the time dimension, that human experts may have traditionally used as a discriminating factor.

In 1964 a flagship study by Mosteller and Wallace marked the first empirical attempt to attribute authorship of the Federalist Papers through the application of statistical discrimination methods (Holmes, 1998). In their study, prepositions, conjunctions, and articles were used as stylistic markers. The frequencies of their occurrence were analyzed and adjusted using Bayesian statistical analysis. "Mosteller and Wallace's scholarly analysis was to open the way to the modern, computerized age of stylometry, and their

work has become a ground-breaking moment in literary detection" (Holmes, 1998, 112).
The computational-stylometric paradigm can be further divided into the similarity-based
(Koppel et al, 2012) or statistical methods (Nirkhi et al, 2015) and the machine-learning-
based approaches (Domingos, 2012). The former measures the similarity/distance be-
tween the texts, whereas the latter applies machine learning techniques to automatically
cluster or classify texts according to labeled training examples.

2.3 Style markers

In the three decades succeeding Mosteller and Wallace's study, much of the stylometry
research has focused on style quantification (Stamatatos, 2009), which yielded a list of
approximately 1,000 stylistic markers (Brocardo et al, 2013; Rudman, 1997). The stylis-
tic features may be organized into 5 feature groups (Stamatatos, 2009), although some
variance in the taxonomy exists (Brocardo et al, 2014; Canales et al, 2011; Roffo et al,
2013). This thesis employs feature taxonomy described by Stamatatos (2009), depicted
in Table 3.1.

A large variety of available stylistic markers poses a dilemma, in terms of which mark-
ers to use? Bailey (1979) asserted that features should be salient, structural, frequent, and
easily quantifiable, but not easily manipulated by the writer (as cited in Holmes, 1998).
Some have noted that feature selection should be based on frequency, as more frequent
features offer a greater stylistic variation (Stamatatos, 2009). Others have argued that the
single features in themselves may not be indicative of authorial style, but may have a
synergistic effect when used together with other quantifiable measures of style (Rudman,
1997).

Table 2.1: Stylistic marker categories

Category	Example
Lexical	Sentence Length
	Vocabulary Richness
	Word Length
	Word Frequency
	Errors
	Word N-grams
	Function words
Character	Uppercase Characters
	Special Characters
	Character N-grams
Syntactic	Part-of-Speech
	Phrase structure
	Sentence structure
	Errors
Semantic	Synonyms
	Semantic dependencies
Application specific	Structural
	Content specific
	Language specific

Feature selection may be performed by human experts or passed on to an algorithm (Stamatatos, 2009). Although the literature presents a wide variety of feature sets (El Bouanani and Kassou, 2014; Forsyth and Holmes, 1996), the jury is still out on the question of the most effective markers (Brocardo et al, 2013). Furthermore, one's writing style may change with experience; the phenomenon termed "drift in style or author-drift" (Arun et al, 2009) which adds another challenge to the already complicated task of feature selection. Nevertheless, the performance of much of the attribution studies ranges between 70% and 100% (Monaco et al, 2013) which suggests considerable potential for mapping learner identities with academic work. The performance is discussed in a separate section.

The n-gram based feature sets are a popular choice and they come in a variety of feature representations (Brocardo et al, 2013; Ding et al, 2015; Howedi and Mohd, 2014; Sidorov et al, 2014; Solorio et al, 2011). For example, Sidorov et al (2014) proposed a feature set based on syntactic n-grams (sn-grams) and compared its performance to character and word based n-grams; Brocardo et al (2013) employed a character 5-gram feature set; Ding et al (2015) employed a feature set comprised of 2,302 stylometric features where 2,000 were of n-grams. Others employed writeprint feature sets (Afroz et al, 2012; Stuart et al, 2013), a combination of function words, parts-of-speech and 1,000 content-based frequent words (Argamon et al, 2009), and an automatic algorithm-based feature selection (Schmid et al, 2015).

2.4 Machine learning

Along with research on stylistic markers, computational techniques have also been a focus of research (Tschuggnall and Gunter, 2014). In 1998, Holmes noted that "the role of artificial intelligence techniques in stylometry seems one of vast potential" (Holmes, 1998, 115), and in the year 2012 Brennan et al (2012) asserted that artificial intelligence techniques have become a dominant approach in stylometric research. Stylometric analysis may be considered a pattern recognition task (Holmes, 1998) which requires a robust computational method. Machine learning may offer just that. Machine learning algorithms recognize the patterns in data and fit the model into new data, with efficiency superior to that of the manual methods. Machine learning may also help to optimize feature selection. Such an approach is particularly beneficial where there is a need to find the best possible algorithm among thousands of available options (Domingos, 2012).

Ayodele (2010) presented a taxonomy of machine learning algorithms comprised of 6 types: supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, transduction, and learning to learn. Much of the authorship studies are conducted using supervised-learning on a closed set of authors (Stamatatos, 2009). That is, a classifier is trained on a set of candidate authors' texts, where the author in question is one of the candidate authors. When learning is supervised, a model is built from a set of labeled data $(x_i, y_i), i = 1...n$ comprised of a feature set X with a class label Y, to learn a function $y = f(x)$ to predict the class label Y for any new values of X. When learning is unsupervised, data are unlabeled $(x_1...x_n)$, and the objective is to organize them into groups sharing similar properties. Authorship attribution problem is generally posited as a classification task (Arun et al, 2009), where each authorial style

represents a class. Classification is "the most mature and widely used" machine learning task (Domingos, 2012, 1). There are seven common algorithm types for conducting classification tasks. These include: Linear classifiers (Logical Regression, Naïve Bayes, Support Vector Machine), Quadratic classifiers, K-Means clustering, Boosting, Decision Tree, Bayesian networks and Neural Networks (Ayodele, 2010). A detailed comparison of classification techniques and description of algorithms is provided by (Cheng et al, 2011). In a real-world setting, the researcher may not always be in control of the quality and quantity of the textual data. Such may be the case with analysis of an anonymous blog posting for the purposes of forensic investigation, where a list of candidate authors is not definitive— as in the open-set problem, where the true author may be unknown— and the data is scarce. To address these challenges the results can be expressed in terms of probability.

Stamatatos (2009) maintained that there are two approaches for passing on training data to the classifier to build a classification model. In the profile-based approach, there is a single document comprised of multiple writings by a single author. This approach treats a sum of different writings as a whole. In the instance-based approach, the training set is comprised of multiple texts stored individually, and divided into blocks of certain character or word lengths. "The majority of the modern authorship identification approaches considers each training text sample as a unit that contributes separately to the attribution model"(Stamatatos, 2009, 16). The case of unsupervised learning takes a more autonomous approach, where an algorithm delineates each class sharing similar attributes within the dataset.

2.4.1 Algorithms

Much of the machine learning algorithms can be applied to textual data; natural language is represented by tokens and converted into numerical vectors. Some of the popular algorithm types used for text classification include : Naive Bayesian classifiers (NB) (Ali et al, 2011; Howedi and Mohd, 2014; Sidorov et al, 2014; Argamon et al, 2009; Cheng et al, 2011); Support Vector Machines (SVM) (Afroz et al, 2012; Brocardo et al, 2015, 2014; Cheng et al, 2011; Ding et al, 2015; Howedi and Mohd, 2014; Pavelec et al, 2009; Sidorov et al, 2014; Solorio et al, 2011; Stuart et al, 2013); Neural Networks (NN) (Pateriya et al, 2014); Decision Trees (DTs) (Afroz et al, 2012; Ding et al, 2015; Sidorov et al, 2014); k-Nearest Neighbors (KNN) (Wan et al, 2012). A comparative study conducted by Ali et al (2011) examined the performance of 14 classifiers including: JW Cross Entropy, KS, Camberra, Cosine, Histogram, Manhattan, Kullback Leibler, Levenshtein, Intersection, LDA, RN Cross Entropy, Naïve Bayes, and Mean Distance. A study by Schmid et al (2015) examined the performance of 4 algorithms which include: AuthorMiner (AM), Classification by Association (CBA), Multiple Association Rule (CMAR), and CMAR for Authorship Attribution (CMARAA). A study by Pavelec et al (2009) compared the performance of compression based algorithm Prediction by Partial Matching (PPM) to that of the SVM. Ensemble methods that combine several different algorithms can also be employed (Cheng et al, 2011), and often yield a higher performance than single classifiers.

2.5 Tools

The previous sections discussed the concept of stylistic markers and machine learning principles for finding patterns in textual data. This section provides a general overview of the software tools that could be used to conduct stylometric analysis.

Machine learning software is used for analysis of a variety of data types, textual data being one of many. Some of the tools however, are designed specifically for textual analysis, and provide built-in domain specific functionality such as feature extraction, tokenization and term frequency-inverse document frequency (TF-IDF) statistic (Ramos, 2003). Much of the software provides an out of the box set of mature algorithms for conducting classification and regression tasks. Some software offers extended functionality to prototype new algorithms or improve on the existing ones.

There are at least three software tools developed specifically for authorial analysis which include: Automated Linguistic Identification and Assessment System (ALIAS)[1]; The Java Graphical Authorship Attribution Program (JGAAP)[2]; and authorship attribution framework JStylo[3]. The former is a commercial software suite comprised of applications for forensic textual analysis developed by ALIAS Technology LLC. (Chaski, 2012). The latter was developed by Duquesne University Evaluating Variation in Language Laboratory (Juola, 2006), while the last is a framework for authorship attribution developed by the Drexel University (Fifield et al, 2015).

`Waikato Environment for Knowledge Analysis (WEKA)`[4] developed at the University of Waikato, New Zealand is an open source machine learning software,

---

[1] http://www.aliastechnology.com

[2] http://www.jgaap.com

[3] https://psal.cs.drexel.edu/index.php/JStylo-Anonymouth

[4] www.cs.waikato.ac.nz/ml/weka/

released under the GNU license. It may be used to conduct a vast array of analyses on a wide array of data types. WEKA was used in conjunction with JavaMail API (Brocardo et al, 2015, 2014) as well as OpenNLP and libSVM libraries (Ding et al, 2015; Pavelec et al, 2009).

`MATLAB`[5] is a language and environment for data analysis and visualization. It is a commercial software developed by MathWorks in the United States. It has symbolic computation capability, making it versatile in prototyping algorithms. MATLAB, was used in combination with C language (Pateriya et al, 2014) and also in combination with Python language (Cheng et al, 2011).

`R`[6] is an open source, released under the GNU license, language and environment for data analysis, predictive modeling and visualization. It is developed and maintained by the R-project core team. It is an open source successor, of the S statistical package. `Stylo R package`[7] is a collection of scripts for stylometric analysis that can be run from within the R console.

`Octave`[8] is another open source GNU license software, both language and environment for computation and visualization predictive modeling.

`Scikit-learn`[9] is an open source, released under the BSD License, data analysis suite based on scientific Python packages NumPy, SciPy, and matplotlib. The software comes with a set of mature algorithms and extensive documentation.

`RapidMiner`[10] is a machine learning framework which comes as both server and desktop versions. It provides an out of the box, modular data analysis environment.

---

[5] http://www.mathworks.com/products/matlab/

[6] https://www.r-project.org/

[7] https://sites.google.com/site/computationalstylistics/stylo

[8] https://www.gnu.org/software/octave/

[9] http://scikit-learn.org

[10] https://rapidminer.com/

The text processing plug-in needs to be installed in order to access the tools for dataset importing and transformation. The software has different licensing options, depending on the version and is available as open-source or as commercial distribution. Rapid-miner platform was used in a study by Howedi and Mohd (2014).

2.6 Datasets

This section provides an overview of the corpora used in stylometric research. The validity of stylometric experiments entails ensuring that analyses are conducted using the ground-truth data (Chaski, 2012). As well, when an algorithm is put to the test, it is imperative to be able to compare the results against the known standards. To this end, a benchmark corpora is developed and provides the necessary reference point for gauging the performance of stylometric techniques. "A proper comparison ... would involve standardized texts of clear provenance, known authorship, on strictly controlled topics, so that the performance of each technique can be measured in a fair and accurate way" (Juola and Baayen, 2005, 2). Stamatatos (2009) noted the trend towards the use of standardized benchmark corpora in authorship attribution experiments. Such an approach is a step towards an objective evaluation of stylometric methodologies, as it attempts to keep the corpora consistent across the studies. He also maintained that it is critical to expand its variety and include different genres and languages. However, there is some disagreement as to what should be used as experimental datasets as not all corpora are of equal reliability and validity. For example, much of the authorship studies are conducted using literary works and some have maintained that the Federalist Papers corpus is a suitable choice for stylometric research namely due to its size and familiarity to the research commu-

nity (Jockers and Witten, 2010). The corpus was used in a number of studies, including the 1964 study by Mosteller and Wallace which some scholars consider as "possibly the best candidate for an accepted benchmark in stylometry" (Forsyth and Holmes, 1996, 20). Others have expressed concerns regarding the validity of experiments using literary datasets, particularly those embroiled in controversy. If the authorship of a literary piece is disputed or unknown, as may be the case with the Federalist Papers corpus, accurate attribution of authorship may not be possible (Stamatatos, 2009, 2).

Another criticism of experiments using literary datasets stems from the notion that literary data are not congruent with the real-world written communications and therefore is lacking ecological validity, limiting its application to the domain of literary research. This may be the case with any content that is undergoing an editorial review. For example, "[t]here is a general worry with newspapers that the texts of the authors are often changed by editor(s)" (Luyckx and Daelemans, 2005, 158). Literary works are formatted, often edited and larger in size than content used in natural communications which are often short, "messy, ungrammatical, unedited, cross-genre, cross-register and sparse" (Chaski, 2012, 337). This raises questions about the nature of corpora to be used in stylometric research and their standardization requirements.Much of the literature does not provide sufficient details for study replication, focusing attention primarily on the results rather than the method, and few studies provide details on data pre-processing methods. However, there are positive attempts to bring scientific rigor to stylometric research. For example PAN@CLEF conference (Stamatatos et al, 2015) provides a platform to comparatively assess and identify the features and computational techniques that yield the best results in each of the analysis tasks. This research strategy is beneficial as it seeks to assert control over the data and enables comparative analyses to be made. Although,

corpora may be compiled from public sources (Stein et al, 2009; Zhao and Zobel, 2007a) and include literary works (Stamatatos et al, 2014; Stein et al, 2011), the competition results allow for a cross-comparison of the proposed solutions applied to the standardized problem sets.

Since 1992, the Text Retrieval Conference (TREC)[11] jointly supported by the National Institute of Standards and Technology (NIST) and the U.S. Department of Defense have been hosting various research tracks to stimulate research of text retrieval techniques. Research questions and related experimental corpora are provided by the NIST. One study Zhao and Zobel (2007a) used the Associated Press newswire dataset from the TREC collection to conduct an authorship attribution experiment. In the year 2004 an Ad-Hoc Authorship attribution competition was held as a part of the Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities, to conduct a series of comparative stylometric experiments (Juola, 2006). The aim was to "establish a collection of the best techniques and methods in inferring document authorship from participants around the world" (Juola, 2004). A few years later, the 2009 PAN evaluation lab[12] on uncovering plagiarism, authorship, and social software misuse began to promote research in the areas of plagiarism detection, author identification and author profiling.

The abundance of electronic textual content, as well as the social nature of the internet provides a favorable environment for the creation and dissemination of experimental datasets. Project Gutenberg[13] currently offers over 49,000 free literary works. Select writings from the project were examined in a study by Zhao and Zobel (2007b). Cor-

---

[11] http://trec.nist.gov
[12] http://pan.webis.de
[13] https://www.gutenberg.org

pus linguistics libraries such as the Linguistic Data Consortium[14] and the International corpus of English[15], offer a variety of texts that are suitable for stylometric analysis. Furthermore, various revisions of the Enron[16] e-mail dataset became available for textual analysis, after the United States Federal Energy Regulatory Commission released the company's e-mail data to the public domain in 2003. The Enron e-mail corpus data is a popular choice of experimental data for stylometric research (Allison and Guthrie, 2008; Brocardo et al, 2014; Johnson and Wright, 2014). A study by Afroz et al (2014) examined the extent of duplicate account usage in four underground forums using an unsupervised learning algorithm that cross-compared authorship of user postings. The dataset was publicly posted to the internet by an unknown entity, in the form of an SQL dump containing user registration, as well as public and private communication. Other datasets comprised of the publicly available information include the IMDb62 dataset which contains 62,000 movie reviews by 62 users of the Internet Movie Database[17] (Seroussi et al, 2011); the Reuter C50 dataset (Nirkhi et al, 2015; Stamatatos, 2007); and Amazon Commerce reviews data set (Liu et al, 2011). The last two datasets are hosted by the Machine Learning Repository at University of California Irvine[18].

New research directions may require the development of a new types of datasets, previously unavailable. This was the case with research on adversarial stylometry and author profiling. Adversarial stylometry is a new area of research that deals with intentional modification of the writing style. In one study, Brennan et al (2012) have built two datasets titled Brennan Greenstadt Adversarial Stylometry Corpus and the Extended-

---

[14] https://www.ldc.upenn.edu

[15] http://www.ice-corpora.net/ice

[16] https://www.cs.cmu.edu/~/enron

[17] http://www.imdb.com/interfaces

[18] https://archive.ics.uci.edu/ml/datasets.html

Brennan- Greenstadt Corpus, comprised of unaltered and obfuscated texts. This corpus was subsequently examined by Stuart et al (2013) in a partial replication study using the non-obfuscation portion of the Extended Brennan-Greenstadt corpus. The author profiling task requires a special dataset comprising textual features and writers' personal attributes. At least two studies examined the link between Jung's personality typology and stylistic features. In a study by Luyckx and Daelemans (2008b) participants contributed their essays to the Personae corpus and also took the Myers-Briggs Type Indicator (MBTI) personality test. The corpus was later used in a study by Noecker et al (2013) who reported a slightly higher results than that of original experiment by Luyckx and Daelemans (2008b).

Similar to literary works, e-mail-based datasets are a popular choice for experimental corpora. For example, datasets derived from the Enron e-mail corpus (Brocardo et al, 2013, 2015, 2014; Cheng et al, 2011; Ding et al, 2015; Schmid et al, 2015) were used in a variety of authorship analysis tasks. Ding et al (2015) used a subset with 2 authors, 20-120 e-mails and 1-320 words per e-mail, whereas a subset employed by Brocardo et al (2015) examined 76 authors with 500 characters, and 50 blocks per author. This dataset is similar to that of two other studies by Brocardo et al (2014) using a pool of 76 authors, 500 characters, 50 blocks per author and by Brocardo et al (2013) with 87 authors with 500 characters and 50 blocks per author. Social media is another popular source of data and includes: the Twitter corpus (Brocardo et al, 2015); the select postings of Chronicle of Higher Education, discussion forum (Solorio et al, 2011); a custom dataset of blog postings by 19,320 authors (Argamon et al, 2009); and the Thomas-Amina Hoax dataset (Afroz et al, 2012). Datasets vary in the number of authors and the size of texts. For example, Argamon et al (2009) employed a dataset with 19,320 authors and in another

study a dataset based on the Thomas-Amina Hoax corpus contained data on 2 candidate authors (Afroz et al, 2012). This often has an impact on performance, which will be discussed in the next section.

2.7 Performance evaluation

It is critical to know how well one stylometric technique performs relative to others, and whether or not the results are sound. The performance of stylometric techniques is affected by several factors including, the type of algorithm employed, the type and number of stylistic features, the number of candidate authors, and corpus size, just to name a few. The study design may lead to bias, for example contamination of test data with training data may affect performance and result in overestimation (Domingos, 2012). The concern about methodological bias has been shared by other scholars (Brocardo et al, 2013; Luyckx and Daelemans, 2008a). "A lot of the research in authorship attribution is performed on a small set of authors and unrealistic sizes of data, which is an artificial situation. Most of these studies not only overestimate the performance of their system, but also the importance of linguistic features in experiments discriminating between only two or a small number of authors" (Luyckx and Daelemans, 2008a, 518).

   The reporting format and the type of metrics used to summarize the results have also been brought under scrutiny. Classification performance may be expressed using a variety of metrics depending on the type of the classification problem at hand: Accuracy, true match rate, precision and recall, and F1 score (Stein et al, 2011) (also F measure or F score is the harmonic mean of precision and recall scores) are some of the common performance statistics reported in the literature. It may also be expressed in terms of

resource consumption such as computing power and time required to complete the task. Much of the studies report performance in terms of the true match rate, that is the number of correctly classified cases from a dataset (Brocardo et al, 2013). The Accuracy score can be expressed as the number of correct predictions (true positives [TP] true +negatives [TN]) divided by the total number of predictions (true positives [TP], true negatives [TN], false positives [FP], and false negatives [FN]):

$$ACC = \frac{TP+TN}{FP+FN+TP+TN}$$

For a detailed analysis of 24 performance measures used in Machine Learning classification tasks please refer to Sokolova and Lapalme (2009).

There are many factors that could influence classification performance such as dataset type and size, corpus language, noise removal techniques, tokenization techniques, normalization methods, the feature set, the number of candidate authors, train/test ratio, classification algorithm type, and its programmatic implementation. Algorithms are also sensitive to noise and scaling techniques. The literature shows that performance results vary across studies, and so do the number of authors, and performance measures. Table 2.2. summarizes a relevant sample of 13 studies.

2.8 Application of stylometry to academic integrity

The rationale for using stylometric techniques to provide identity and authorship assurance in the learning environment is that student-generated content is readily available and carries individual-specific patterns. Students employ perceptual filters and given the same

Table 2.2: Summary of performance rates and measures

Reference	#Authors	Performance	Measure
Afroz et al (2014)	12; 35; 2	9.5 %-98 %	F-Measure
Argamon et al (2009)	19320; 1290; 198	20.0 % - 82.3 %	Accuracy
Brocardo et al (2013)	87	14.35 %	EER
Brocardo et al (2015)	76; 100	9.98 %- 21.45 %	EER
Brocardo et al (2014)	76	12.42 %	EER
Ding et al (2015)	2-20	30 % $\leq$90 %	Accuracy
Howedi and Mohd (2014)	10	20.00 %-96.67 %	Accuracy
Schmid et al (2015)	10	15.78 %-86.9 %	Accuracy
Sidorov et al (2014)	3	33 %-100 %	Accuracy
Solorio et al (2011)	5-100	32.77 %-77.38 %	Accuracy
Stuart et al (2013)	5-40	94 %-100 %	Precision
Pateriya et al (2014)	2-8	65.39 %-98.65 %	Accuracy
Pavelec et al (2009)	20	83.3 %	Accuracy

input, different students respond differently. The comprehension of reality is conducted through the lens of existing beliefs and assumptions that impose limitations on how the perceived inputs are construed. This results in the production of academic artifacts containing a distinct signature, particular to each student. Therefore, artifacts produced by the same student are expected to be more similar to each other than to the work of other students. This allows a delineation of student-produced artifacts by analyzing stylistic choices exercised by students. These artifacts are then collected and stored for analysis,

bearing a label of authorship that allows subsequent identification. The process of content creation is depicted in Figure 2.1.
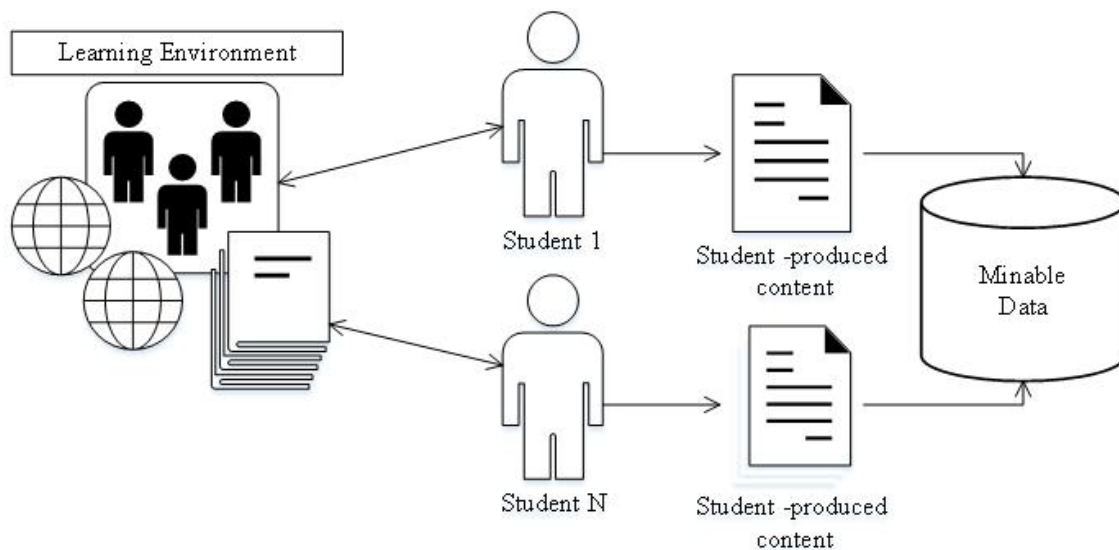

Fig. 2.1: Content creation process

Because learning is a continuous process, students produce content as they progress through learning activities. Machine learning provides the necessary means to analyze the data. Prior academic work can be used to create a stylistic profile that will be tested against all subsequent student-produced content. Even before program enrollment, schools often require their prospective students to complete entrance exams, which may be used as inputs to the validation process of subsequent learning activities. The process of analyzing student assignments is depicted in Figure 2.2.

The problem of aligning student identities with the work they do is posed as a classification task. Given a set of documents, an algorithm associates textual features with the labels that represent identities of the students who produced them. When a new artifact is
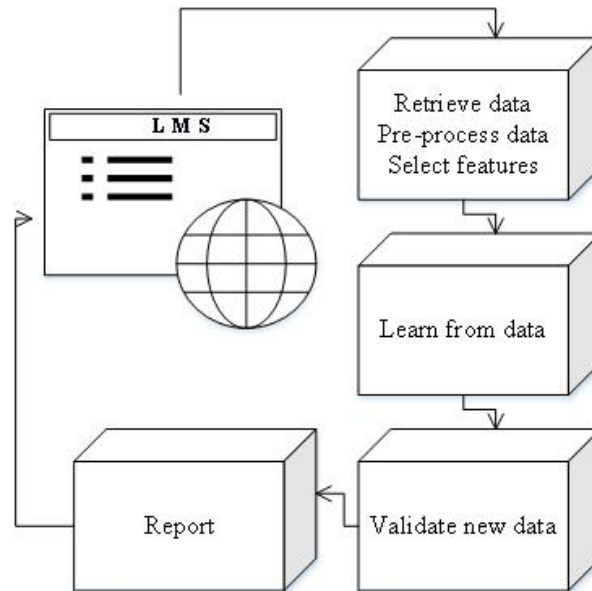
Fig. 2.2: Data analysis of student-produced content

presented, the algorithm predicts a student whose use of textual features is more similar to the ones learned earlier. The student-generated content is passed on to classification algorithm(s) that learn to associate labels (student names) and patterns of language use from the examples in the training set and predict a class label which represents the students for each sample in the testing set. This prediction is then compared to the student names at the time of the assignment submission and any discrepancy in the predicted labels versus the student-supplied labels raises a red flag. To cover any blind spots in the academic environment, any cases of misclassification should be randomly examined by the instructor to ensure that the standards of academic integrity are maintained.

Figure 2.3 depicts two classification scenarios. Scenario A features an artifact claimed by Student X that was classified to be produced by Student X, and an artifact claimed by Student Y that was predicted to belong to Student Y. In contrast, Scenario B depicts

a case of misclassification in which an artifact produced by Student Y bears similarity to the stylistic profile of Student X, in spite of being claimed by Y, which suggests a conflict and calls for the instructor's attention.
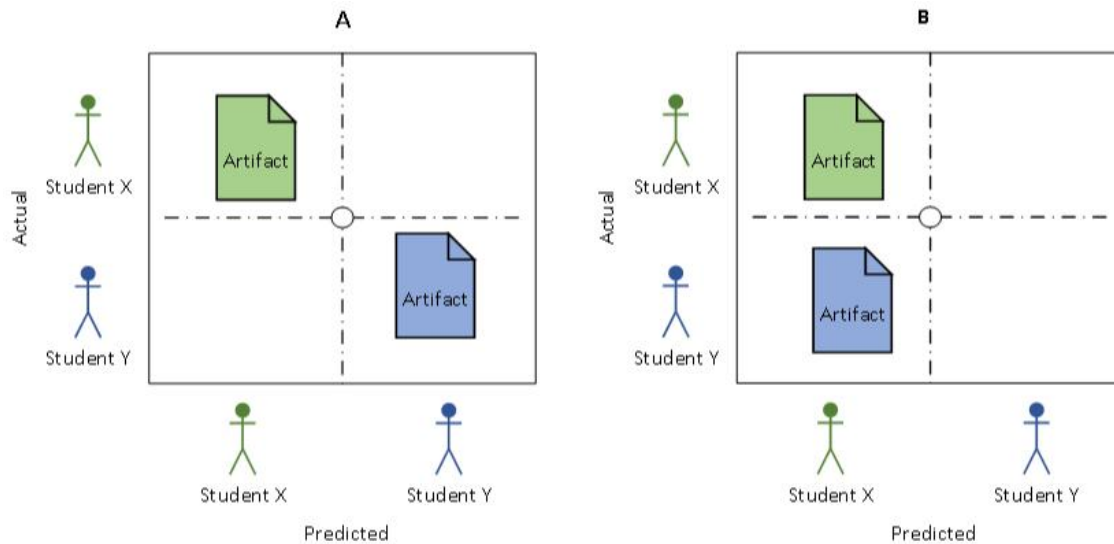
Fig. 2.3: Classification scheme

2.9 Summary

This chapter explored the notion of authorial discrimination, provided an overview of methods and tools used in authorship analysis tasks, and discussed the application of stylometry to the issue of academic integrity. It also provided examples of machine learning algorithms that were applied to the analysis of textual data and examples of experimental datasets. Much of the attribution studies show performance rates in the 70% to 100% range (Monaco et al, 2013) , which suggests that stylometry may open the path to a

one-tiered identity and authorship validation of academic content. However, the literature varies widely on the analysis techniques and datasets. There is a general agreement that factors, such as the number of candidate authors, the size of the training and testing sets affect performance (Ding et al, 2015; Luyckx and Daelemans, 2008a; Stamatatos, 2009); however, there is no one-size-fits-all approach available, and the search for the most effective techniques is still ongoing.

The following chapters propose, develop, and validate an identity and authorship assurance framework that uses a combination of stylometric and machine learning techniques for mapping of learner-produced content to their identities.

Part II

Empirical Perspective

This part comprises two chapters devoted to the development and evaluation of the behavioral-biometrics-based and computationally-driven approach to academic integrity. The analyses were conducted using a set of real-world academic writings, which are representative of the type of artifacts the students are expected to produce. It commences with a presentation of the proof of concept in Chapter 3, then a refinement to the method using ensemble techniques is introduced in Chapter 4. The performance of the computational approach and its comparison to the efficacy of the instructor validation approach— used as a baseline— are discussed. The results suggest that the computational approach outperforms the instructors in classifying short texts by author.

Chapter 3

Conceptual Framework and Proof of Concept

Overview

This chapter introduces an academic integrity framework for e-assessment that provides identity and authorship assurance through pattern analysis of the student-produced content. The proposed framework utilizes stylometric and machine learning techniques that analyze patterns of language use, providing a concurrent, accessible, and non-disruptive validation of student identities and student-produced content. This approach attempts to maximize convenience for the learner and lower administrative overhead for institutions, stemming from the reduced need for managing physical space and human resources. It also aims to empower instructors with automated tools that promote accountability and academic integrity, while providing accessible and non-invasive validation of student work.

The challenge with the provision of secure assessment in a distance environment is due to two overlapping tests that need to be conducted for ensuring integrity during the assessment activities: first, the identity confirmation test, and second, the authorship validation test. The former is concerned with verifying that students are who they claim they are, while the latter is concerned with validating that students do what they claim they do.

To address this challenge, many of the institutions opted to seek alternative assessment arrangements (Bailie and Jortberg, 2008) and conduct assessments external to the LMS (see Figure 3.1).
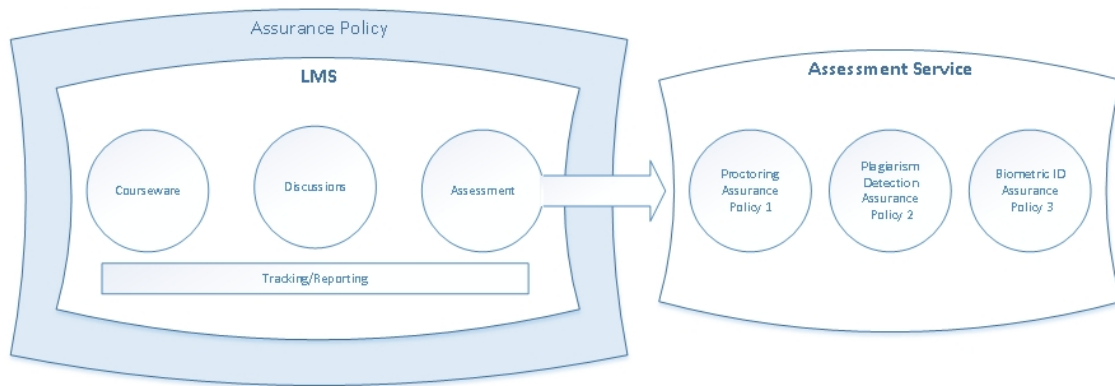
Fig. 3.1: Externalization of assessment activities

3.1 Proposed framework

The proposed method examines learners' stylistic preferences, measuring and monitoring the production of the written academic content throughout the course life-cycle. In contrast, the traditional approach to secure assessment is based on observation and is often employed in only high stakes examinations, because underlying administrative and logistical overheads make it costly and challenging to monitor all of the learning activities. Figure 3.2 depicts the difference in approaches. The proposed approach has several advantages over the traditional one. It does not require any special hardware, and can be integrated with the LMS for automatic data collection and analysis. It provides concurrent identity and authorship assurance, where a written data is used for identity and
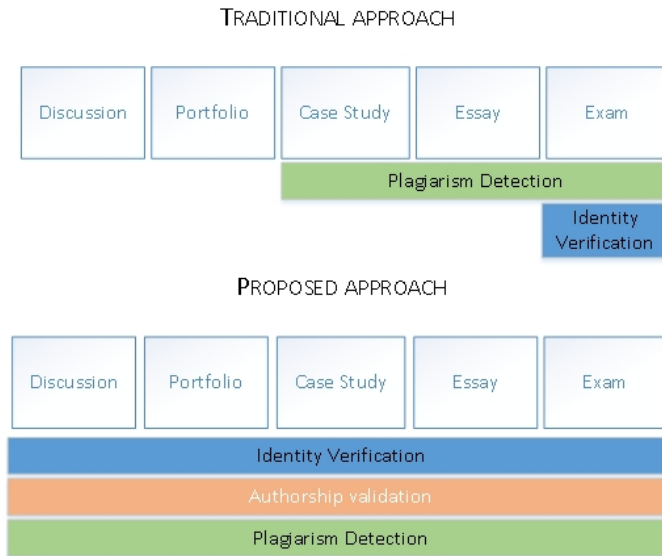
Fig. 3.2: E-assessments assurance approaches

authorship verification. This entails lower scheduling requirements, lower administrative overhead and a consistent user experience across learning activities.

However, the method has limitations. It is designed to work with authentic assessments such as essays, projects, and portfolios. Validation of multiple choice tests is outside of its current scope. As with any biometric approach, it is prone to type I and type II errors. The method is not designed to completely replace instructor validation, but rather to simply the task of aligning learner identities with the work they do.

3.2 Implementation

The proposed framework is comprised of four distinct functions: enrollment; analytic task creation; content collection; analysis and reporting. Similar to other identity control

applications, the process commences with an enrollment. Learners provide a copy of a writing sample to be used for training the classifier. At this stage, a stylistic profile is created and will serve as baseline for future comparisons. Instructional designers or instructors incorporate assessment activities into the course content. Some of the activities generate written content, and these activities may be defined as identity and authorship assurance tasks, where the learner-generated content is analyzed against that in their stylistic profile. The end result of this analysis is a report provided to the course instructor flagging cases that require intervention. For large course sizes, a random sample may be drawn to validate select cases. The process workflow is depicted in Figure 3.3. All assessment activities and their subsequent verification is performed within the LMS, either as a module or a stand-alone application with integrated user management. Students are not required to contact third-party providers, acquire additional hardware and schedule assessments. To the contrary, the process is minimally invasive for both the learner and instructor, allowing them to focus on learning and teaching.

The proposed approach is different and has distinct advantages in that:

- Stylometry behavioral biometric is used for both identity verification and authorship validation.

- All student-generated textual content is validated continuously.

- The process of validation is non-invasive and does not require learner interaction. Frequent identification requests may affect the level of perceived convenience (Ullah et al, 2012).

- Assessment activities are conducted within the same learning space as the course content, promoting accessibility through consistent user experience and familiar environment.
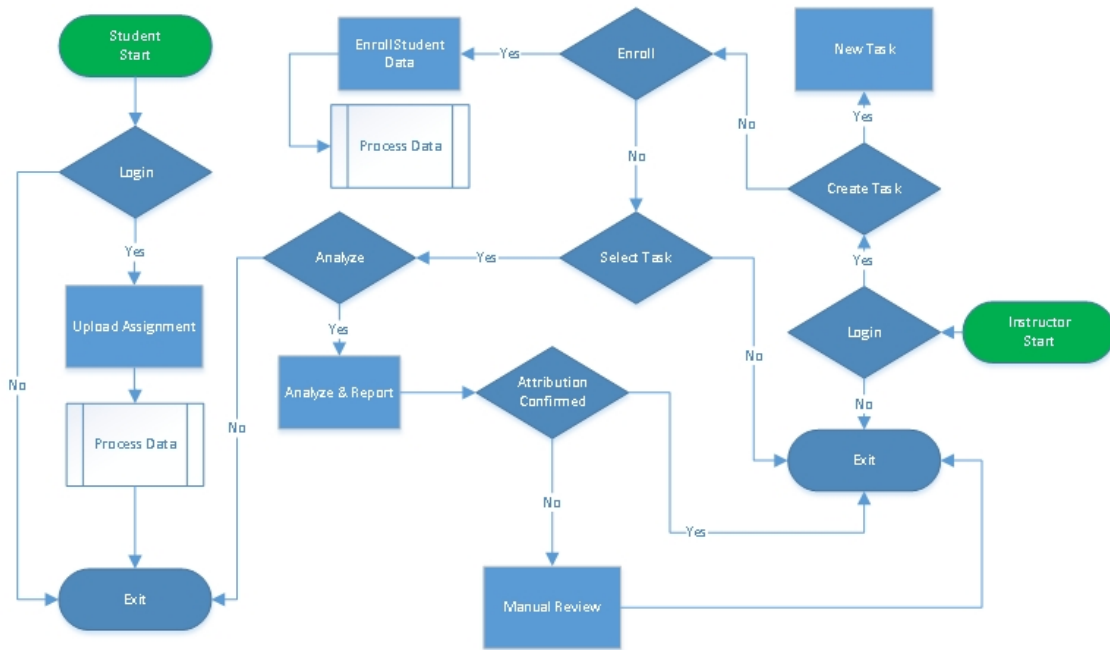
Fig. 3.3: Proposed framework

- Behavioral biometric data is maintained by the academic institution, yielding a greater degree of buy-in from students. Learners are more likely to use biometric authentication through their university than that managed by a third-party provider (Levy et al, 2011).

3.2.1 Enrollment

The academic application process entails a set of verification activities to meet program entrance requirements. Prospective students provide personally identifiable information (identity proofing) such as government issued identity documents, original transcripts, letters of reference, proof of language proficiency, and often samples of their work. Upon

acceptance into the program, applicants' status changes to that of a registered student. The academic services department provides new students with the means of access to academic, administrative and its IT services by creating an identity profile (identity enrollment) and issuing credentials such as a student identity card and user credentials. More often than not, user credentials are in the form of a student number or institutional email and password pair. User credentials are then used as a digital representation of students' physical identity. User credentials are comprised of factors that fall under the four categories which include: something one knows (e.g., student number or password); something one has (e.g., mobile phone or identity card); something one is (e.g., fingerprint or retina image); or something one does (e.g. speech patterns, writing style). All subsequent course enrollments and access to academic resources are conducted using the established credentials such as student number and password pair. Although the student profile may contain information on their academic abilities and stylistic peculiarities, this information is not being reviewed beyond the decision for program enrollment. The proposed approach takes these data into account to establish an additional authentication factor, that is what the learner does. Stylistic preferences are extracted from the prior works such as essays or published articles and the learner stylistic profile is created. If writing samples are not readily available, it will be necessary to create them at the time of enrollment.

After sample writings are collected, the text undergoes pre-processing that includes standardization of encoding, removing title pages, tables of content, direct quotes, symbols, bibliographies, and other noise items. The features are then extracted and may include any combination of the lexical, character, syntactic and application-specific features (Stamatatos, 2009). Vectorization turns feature counts into term-document matrix.

Depending on the type of vectorizer used, term-document matrix may be either sparse or dense. Sparse format is often employed in text classification where the number of features may be large, as only nonzero entries and their positions are stored, which yields space saving.

3.2.2 Analytic task creation and content collection

Each course is comprised of a number of assessment activities that range from weekly forum postings to written assignments to portfolios to timed exams. Each assessment activity may be considered a two-tiered assessment task that includes domain-specific assessment and identity and authorship validation. In the former task, instructors evaluate the quality of the produced content, award grades, and provide feedback on how to improve understanding of the subject; and in the latter task, instructors validate veracity of authorship and verify identities of the students taking part in assessment activities. Each assessment activity deemed important to undergo identity and authorship verification needs to be defined. For the purposes of the framework, such activity is considered an analysis task. This allows for exclusion of any activities that do not generate textual content or do not require identity or authorship validation. Once the analysis tasks are created by the instructor, assignment submission is mapped to each task. Because all submitted content is verified against a global stylistic profile (see Figure 3.5), changing task order will not affect the results. Analysis tasks can be created anytime during the course.

Data collection and processing methods depend on the type of technology used in facilitation of learning activities. For example, a written assignment can be uploaded by

students to the learning management system or external document repository. For these types of activities, file format and document format can be specified by the instructor in advance. Students may be asked to submit assignments only in the portable document format (PDF) where document style does not include headers or footers. In a case where file format and content formatting requirements are not specified, content may be presented in a variety of popular formats and would need to undergo an extraction process specific to the methods by which the textual content was created. Activities such as discussion forums are stored in the LMS database, and again, the method of extraction would be specific to the technology used. All artifacts marked for authorship and identity analysis are then extracted to their specific task space (Figure 3.5), and undergo the same processing steps as the training data. Namely, the encoding is standardized; noise items removed; features extracted and vectorized.

3.2.3 Analysis and reporting

The goal of the analysis task is twofold: first to predict class labels of the student-generated content in each instructor defined learning activity, and second, to compare the prediction to their expected values. That is, all artifacts created by the same student across learning activities are expected have the same class label. At the crux of the proposed framework lay several assumptions: first, that two documents written by the same author share more stylistic features than that written by different authors. Second, both the training and testing datasets are labeled. Third, the training set is considered ground truth data, whereas the test set labels are assumed to be true, unless classification results suggest otherwise. Validation of academic artifacts becomes the process of hypothesis

testing, where H_0=authorship claim is true and H_1=authorship claim is false. Fourth, impersonation, collusion and plagiarism are assumed to affect distribution of the stylistic features. Fifth, the problem of identity and authorship assurance for the purpose of e-assessment can be formulated as a closed set problem (Stamatatos, 2009), whose aim is not to identify the anonymous writer, but rather to confirm or refute the authorship claim made by the student. To this end, the classifier is first trained using the training dataset, and then the model is fit into test data, which outputs class prediction. A confusion matrix is then constructed to visualize classification results (Figure 3.6). Depending on the number of assessment tasks and course design, some steps may be iterative. The size of dataset is proportional to number of assessment activities and their respective size of textual content. The framework is flexible enough to employ an iterative process of feature and algorithm selection, as well as provide interchangeability of computational and stylometric techniques that in addition to author verification and identification tasks may also perform plagiarism or collusion detection and author profiling analyses.

The algorithm depicted in Figure 3.4 summarizes the process of classification and hypothesis testing.

Let D be a corpus of documents $D = (d_1...d_i)$. Let V be corpus vocabulary, comprised of terms $W \in V$. Let $S = \sum_{w_i \in V}$ be the size of corpus vocabulary. Let X denote a feature set with a class label Y, $(x_i, y_i), i = 1...n$. Let each document be a vector of frequency values of occurrence of terms in a document $x = \int \{w_i, d\}_{w_i \in V}$. Let $Y = (y_1...y_i)$ denote class labels of the training set, and $Z = (z_1...z_i)$ denote class labels of the testing set.

Fig. 3.4 Authorial verification in e-assessment

1: **procedure** ENROLLMENT

2: Load training set as (d_i, y_i)

3: **end procedure**

4: **procedure** DATA COLLECTION

5: **For** each task A retrieve testing data **end For**

6: Load testing set as (d_i, z_i)

7: **end procedure**

8: **procedure** PRE-PROCESSING

9: **For** each document d:

10: Standardize encoding

11: Remove noise (eg. title pages, tables of content, direct quotes, bibliographies)

12: **end For**

13: **end procedure**

14: **procedure** BUILD THE MODEL

15: Select classifier

16: Define feature set

17: Tokenize terms w

18: Vectorize tokenized data

19: Learn dictionary

20: Build feature matrix

21: fit (x_i, y_i) into the model

22: **end procedure**

23: **procedure** PREDICT CLASS

24: **For** each x predict label $y = maxP[y_i|x_i]$

25: **end For**

26: **end procedure**

27: **procedure** REPORT AND VISUALIZE

28: **For** each $y \neq z$ flag for manual review

29: **end For**

30: Produce report

31: **end procedure**

3.3 Preliminary data

A series of preliminary experiments was conducted to test the framework on real-world data and to establish a baseline for future research. This proof of concept study aimed to test whether it was possible to verify authorship of learner generated content, and by extension, verify learner identity across multiple learning activities using a bag of n-grams approach (Sapkota et al, 2015).

3.3.1 Data collection

Textual data was obtained from a Research Methods course and included forum messages and essays of 11 students. The course is prerequisite to thesis writing and is aimed at graduate students. The course was conducted in English for non-native speakers. The assessment activities included short group discussions, two assignments on research methods, one research proposal, and one assignment on formal proofs. The last was excluded from data collection, as it involved formal proofs. The topics of the short group discussions included critical thinking exercises such as critique of an article. Assignment one was comprised of two distinct tasks. First, students were asked to answer questions on the subject of research methods, citing the textbook. One example of a question is, "Choose the five most relevant objectives of a literature review." In the second part of the assignment, students were asked to outline a research proposal, that would include a provisional title and a list of articles for the literature review. The proposal outline of part two of Assignment One was later elaborated into a full proposal in the Assignment Four and as such there was a partial overlap between the two assignments. Assignment

Three was similarly comprised of two distinct tasks. First, the question and answer portion, again asked students to cite the textbook to answer the questions such as "Choose the three most relevant advantages and disadvantages of interview-based research." The second part involved designing a mockup survey.

Assignments were submitted in various formats including PDF, Word Document, Latex, and also ZIP archives that contained a combination of documents and graphics. Group discussions were in the EML format (RFC 822 standard). Papers were produced using a variety of templates, that varied in overall presentation of content: citation styles, front page information, headers, footers, page numbers, and bibliographies were inconsistent across students. Much of the student papers reiterated original questions verbatim, with an answer underneath them.

The size and type of the feature and data sets is a limitation of this study.
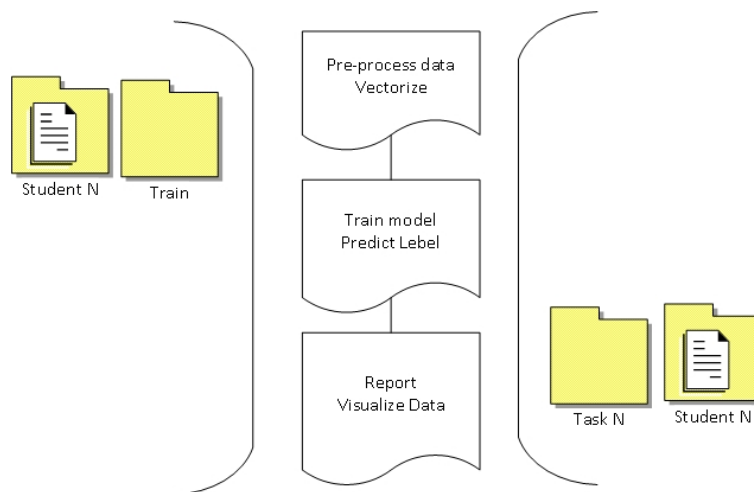
Fig. 3.5: Classification scheme

3.3.2 Method

All data was converted to plain text format and lowercase. Student names, student numbers, title pages, headers, footers, tables of content, bibliographies and stop words were removed. Direct citations and assignment questions were not removed. The document lengths were as follows: Assignment One: 1200 - 5700 words; Assignment Three: 1900 - 4500 words; Assignment Four: 800 - 3600 words; One forum posting of every user on the topic of critical thinking was selected. Much of the forum messages fell in the 250-450 word range. The profile-based approach was employed, using a single document per user and imbalanced training sets (Stamatatos, 2009). The analyses were conducted using scikit-learn (Pedregosa et al, 2011), Multinomial Naïve Bayes classifier (Domingos, 2012), CountVectorizer and word n-grams as style markers (n=2,3). The use of n-grams as style markers has been extensively examined and has shown promising results (Antonia et al, 2013; Brocardo et al, 2013; Sapkota et al, 2015). To measure the performance, the classification accuracy measure was used, expressed as percentage of the number of correct predictions made over the total number of predictions made. The measure of accuracy has limitations; however, the aim was not to find the best algorithm-feature-set combination but rather to test the framework on real world data. Some of the important conditions in performance evaluation include: test and training corpus size, number of candidate authors and whether or not the training corpora is imbalanced (Stamatatos, 2009). These considerations will be reflected in the findings summary. There were five experiments conducted in total and the protocol was as follows:

1. Train classifier on Assignment One to predict authorship of Assignment Three
2. Train classifier on Assignment One to predict authorship of Assignment Four

3. Train classifier on Assignment One to predict authorship of a Forum message

4. Train classifier on Assignment Three to predict authorship of a Forum message

5. Train classifier on Assignment Four to predict authorship of a Forum message

3.3.3 Results

The aim of this study was to examine whether an artifact produced in one learning activity could predict authorship of an artifact produced in another activity and, by extension, confirm the identity of the learner who submitted the work. The task was quite complex, as topic and register were not consistent across all activities. For example, group discussions were colloquial, whereas the assignments were more formal. There was some overlap of topics, not only between the assignments, but also in the use of the textbook that was cited by all students. The data were assumed to be the ground truth and students followed the academic integrity policy.

Table 3.1: Proof of concept results

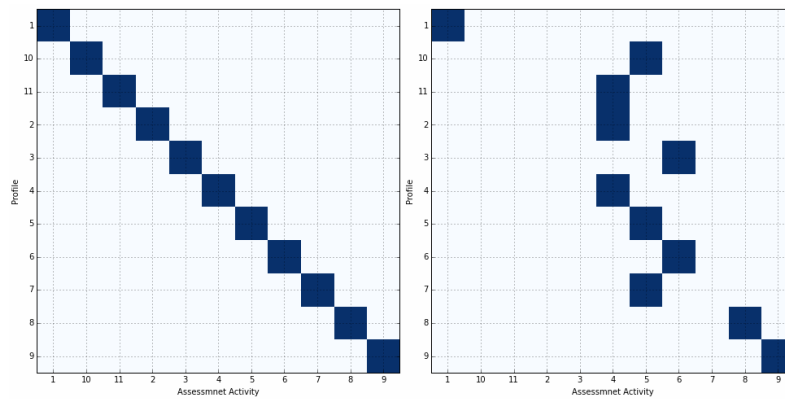| Experiment	Authors	Train Size	Test Size	Accuracy
1	11	1200-5700	1900-4500	54.5%
2	11	1200-5700	800-3600	100%
3	11	1200-5700	250-450	27.3%
4	11	1900-4500	250-450	18%
5	11	800-3600	250-450	54.5%

Fig. 3.6: Confusion matrices (Experiments 2 and 5)

The results, presented in Table 3.1, suggest that topics seem to play a role in variance of performance. Experiment Two yielded 100% accuracy score, which was not incidental, but due to an overlap of language used in both assignments. Figure 3.6 depicts classification results of experiment two (left) and experiment five (right). Assignment Four was an extended version of the proposal outlined in Assignment One. This may also be interpreted as a case of self-plagiarism, because all students have reused word bi-grams and tri-grams from Assignments One in Assignment Four. All of the eleven research proposals were on unique topics. The classification accuracy of Experiment One was 54.5%; that is, the classifier made the correct prediction in just little over half of the cases. Both assignments were comprised of two tasks, a questionnaire portion covering different aspects of research methodology where students answer questions with reference to the textbook; and the second task of the two assignments involving outlining a proposal and developing a mockup survey. In-text citations were not removed and that could have contributed to the noise and misclassification. The results of experiments Three, Four, and Five that aimed to identify authors of the forum postings using Assignments One, Three and Four also varied. The highest accuracy score of 54.5% or about a half of the partic-

ipants was attained using the research proposal as a training set and the lowest score of 18% was attained using Assignment Three as a training set. Here again, document topics seem to be linked to performance. The group discussions were on the topic of critical thinking, and students were asked to critique an article. In Assignment Four learners were asked to review literature, which appears to be thematically closer to group discussion than Assignments One and Three. Training corpus length also appears to play a role in classification accuracy, where the classifier is biased towards larger texts, which was not the case with Experiment Two. The results should be taken as preliminary, as the aim was to test the framework on real data. The data remained noisy with direct citations and assignment questions throughout the text.

3.4 Summary

This chapter discussed a behavioral biometrics-based and machine-learning-aided academic integrity framework for e-assessment that enables non-invasive and continuous analysis of academic content throughout the course life cycle. Unlike the traditional methods that often employ observation, the proposed approach is based on analysis of student-produced content and is expected to alleviate logistical burdens, minimize administrative overheads, and replace human invigilators with computational techniques that facilitate automation of identity and authorship assurance tasks. Preliminary tests were conducted, and the results suggest that the proposed method can be used for aligning learner identities with the work they do, contingent upon establishing a baseline for classification accuracy that is at par or superior to instructor validation. The next steps

would be to improve performance, establish a performance baseline, and compare the ability of human instructors to verify learners by their writing style.

Chapter 4

Human vs Machine

Overview

This chapter expands on the preliminary proof of concept study and presents an improved method for aligning learner identities with their academic work. In contrast to the earlier approach, the new method employs automatic feature selection using the Random Forest Classifier and computing mean decrease impurity (Louppe et al, 2013) and the majority rule voting technique (Raschka, 2015). To assess the efficacy of the computational approach, a baseline of instructor performance— classifying student writings— was established. The results suggest an improvement over the earlier approach, and that the computational methods outperform the human ability to classify texts by authorial style. To the author's knowledge, this work is the first of its kind to compare the performance of instructors to that of computational methods.

4.1 Ensemble methods

Ensemble methods combine output from multiple algorithms which often leads to an improved performance (Cortes et al, 2014). The majority voting algorithm—an ensemble technique— takes class predictions as input and measures central tendency (mode) of the predicted labels by each of the classifiers in ensemble where the most frequent class label wins (Raschka, 2015). Outputs of individual classifiers (each document in the dataset is a member of a single class) are used as inputs to the majority voting classifier that counts the labels returned by each of the classifiers. A majority voting can be expressed as:

$$\hat{y} = mode\{C_1(\mathbf{x}), C_2(\mathbf{x}), \ldots, C_m(\mathbf{x})\}$$

For example, if there are three classifiers in ensemble and two classes, Student A and Student B, and two of the classifiers predict that the assignment belongs to Student A, then the majority wins and the class is predicted as Student A. Figure 4.1 demonstrates this example.

4.2 Proposed approach

Building upon the proof of concept, the new approach aims to improve performance by employing tree-based feature selection technique, non-contiguous word pairs, and ensemble-based classification technique. The feature selection process employed weighting. Many of the studies employ hard feature selection that is, all features are assumed to be of equal importance. However, this approach identifies the top 300 features that are selected. This entails weighing and sorting the extracted features in order of importance.
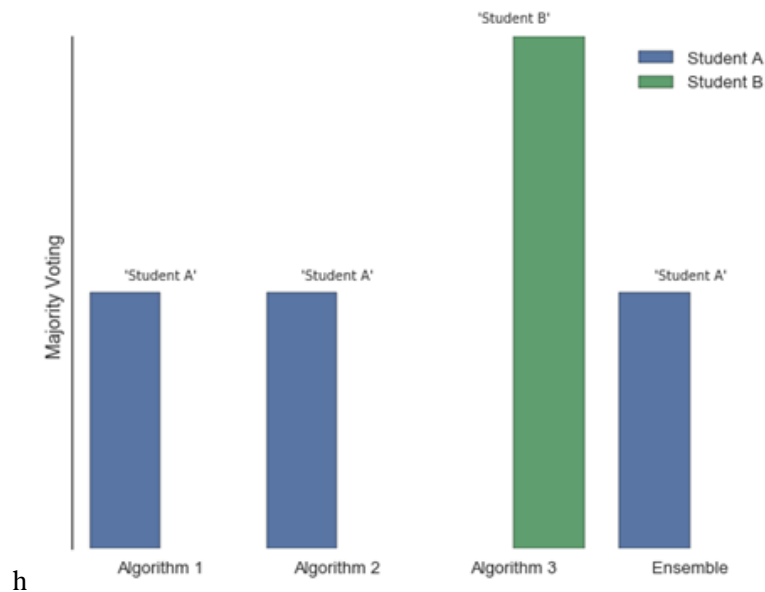
Fig. 4.1: Feature weighting

Feature importance weighting using the Random Forest Classifier and computing mean decrease impurity (Louppe et al, 2013) was performed. Feature weighting is concerned with ranking feature importance, and reducing the dimensionality of the vector space (dropping variables that are redundant or not important.)

The proposed method employs the majority rule voting technique using three state of the art classification algorithms commonly used in text classification tasks. To compare the performance of the new method to that of the proof-of concept, several analyses were conducted. The study protocol was as follows: The raw content was retrieved from the learning management system (LMS) data stores and pre-processed (removing the noise items, tokenizing text and vectorizing tokens, split into chunks); The features were weighted and reduced to the 300 most important features; The dataset was fit by the three classifiers returning class labels for each classifier; The class labels were used as inputs to the majority voting algorithm which returned the predicted class label. A 10-fold cross-

validation technique was used to attain more accurate evaluations. Visualization of class

predictions was performed. Figure 4.2 summarizes the algorithm employed.

Fig. 4.2 Data analysis algorithm

procedure ENROLLMENT
2: Load training set as (d_i, y_i)
end procedure
4: **procedure** DATA COLLECTION
For each task A retrieve testing data **end For**
6: Load testing set as (d_i, z_i)
end procedure
8: **procedure** PRE-PROCESSING
For each document d:
10: Standardize encoding
Remove noise (eg. title pages, tables of content, direct quotes, bibliographies)
12: Split into S word chunks
end For
14: **end procedure**
procedure FEATURE SELECTION
16: Define feature set
Tokenize terms w
18: Vectorize tokenized data
Build feature matrix
20: Weigh feature importances
Select top F features and drop the remainder **end For**
22: **end procedure**
procedure BUILD THE MODEL
24: Select classifier(s) C
For each C fit (x_i, y_i) into the model **end For**
26: **end procedure**
procedure PREDICT CLASS
28: **For** each x predict label y **end For**
end procedure
30: **procedure** WEIGH OUTPUTS
Compute MODE of outputs from C
32: **end procedure**
procedure REPORT AND VISUALIZE
34: Visualize label mapping
Produce report
36: **end procedure**

4.3 Approach evaluation

To assess the performance of the proposed approach a corpus of real world student writ-

ings was obtained. The dataset comprised written assignments enrolled in two graduate-

level Research Methods courses at a fully distance European university. The courses were delivered in English, and all of the students were non-native speakers of English. The courses employed an authentic assessment method. These assessments were low stakes and not proctored. The course integrity was maintained through instructor validation and an academic integrity policy. The collected data were assumed to be the ground truth; that is, the authorship claim for each document was considered to be true and that students were responsible for producing their own work while adhering to the academic standards. Student content was produced using tools and methods that students deemed fit in the circumstances. Students have used a variety of templates and formatting styles to present their work, hence there were differences in the amount of footnotes, headers, footers, in-text citations, bibliographies, and the amount of information on the front page. Student assignments were provided in a variety of formats and included: Word documents, Portable document format (PDF), Latex, and Zip archives containing graphics and documents. Forum postings were originally stored in the EML format (RFC 822 standard). Upon retrieval of the learning management system, all content was converted to plain text (UTF8) format. Extraneous information such as headers, footers, names, e-mail addresses, front pages and bibliographies, such that could identify students directly were removed. Upon retrieval from the LMS datastore, the data were made anonymous and all identifiable information was replaced by a participant number.

Data retrieval and text processing tasks were automated using the Python language. The experiments were performed using Scikit-learn, the python machine learning library (Pedregosa et al, 2011). Performance evaluation was performed using the 10-fold cross-validation method. The same train/test split ratio was used as in (Brocardo et al, 2015; Schmid et al, 2015). For more accurate examinations of the approach, an evaluation was

conducted using the 10-fold cross-validation method, and used the same train/test split ratio as many of the other authorship studies (Brocardo et al, 2015; Schmid et al, 2015). Performance was reported using the accuracy score.

4.3.1 Human performance baseline

The literature is sparse on performance and effectiveness of academic integrity strategies. To bridge this information gap, a study was conducted to establish a baseline of human performance and to measure how well the practitioners directly responsible for grading assignments can classify student writings by author. Barnes and Paris (2013) argued that the instructors should be able to identify instances of cheating or plagiarism once they become familiar with the student's writing. This theory was put to the test. The experimental protocol was as follows:

1. Retrieve student assignments (raw data)
2. Select texts by 5 authors
3. Process data (e.g., remove noise)
4. Create 6 question sets (5+control)
5. Invite participants to perform classification
6. Compute accuracy of predictions

The texts (500 word excerpts) were randomly distributed over five classification tasks. There were three texts per task. Two of the three texts were written by the same student. Five multiple choice questions accompanied each task, asking the participants to identify which texts were written by the same student. One task was used as a control, where all

three texts were the same. Participants were included in the study if: (a) they taught at an accredited university, (b) their professional responsibilities involved grading student assignments in the recent academic year, and (c) they correctly identified that the texts in the control task were by the same author.

The original experiment design had two conditions: (a) the texts preserved original formatting, or (b) the texts were lowercased with punctuation removed. In the latter condition, the dropout rate was high—participants reported difficulty completing the task citing unnatural writing style. As one participant noted: "The elimination of punctuation in the samples makes it very hard to read for stylistic patterning and thus same-author identification." Only the results from the former condition are reported here. Data from completed responses by 23 participants was analyzed. The mean accuracy was calculated as 12% for the group. Some of the markers that the participants used to differentiate authors included: the use of "Britishisms", the use of punctuation, quality of writing, misspellings, "distinct structure", and "unconventional use of language." Although, the human performance data provides a general idea of how well the instructors perform classifying student writings, one would expect the performance to decrease in the real-world setting with an increase in the sample size, text size and type, fatigue, etc. Deviation from the conventional style (e.g., all lowercase text) appear to present a barrier to completion of the classification task. A larger study using a corpus of texts in various genres, reading levels, and topics will be necessary for a more precise evaluation.

In the next sections, the instructor performance rate is used as a baseline for comparison to that of the computational methods.

4.3.2 Evaluation 1

The first evaluation employed the dataset used in the instructor validation study and comprised two-500 word texts by five students. The texts were converted to lowercase and punctuation removed. The feature set was comprised of spanning intervening bigrams (non-contiguous word pairs) with the windows size =10, frequency of occurrence ≥ 2. Function words were preserved. Term Frequency-Inverse Document Frequency (TF-IDF) weights were computed. Some features are more important than others, and isolating them often helps to improve computational efficiency and performance. To this end, the top 100 features were selected using the Extra Trees Classifier. The analysis was performed using the SVC classifier with Linear kernel. Considering the modest sample size, half of the data was used for training and half for testing, randomly shuffled over 10 iterations. This yielded a mean accuracy rate of 92%.

4.3.3 Evaluation 2

The second evaluation employed a dataset comprised of student-generated content by five students enrolled in two graduate-level research methods courses. The corpus was composed of four assignments, two in each course, ranging between 1000 and 6000 words. The documents have undergone pre-processing steps, and noise-contributing items, (e.g., students' names, course numbers, and citations) were removed using a set of regular expression rules, fine-tuned with each iterative step to address specifics of the documents. Documents were split into chunks of 500 words.

The feature set was comprised of spanning intervening bigrams (non-contiguous word pairs) with the windows size =5, frequency of occurrence ≥3 and parts of speech (POS). Function words were preserved. A syntactic set of 41 POS tags was extracted using the NLTK POS tagger; please see The Penn Treebank tag set for more information (Marcus et al, 1993). Feature extraction was performed using the NLTK library (Bird, 2006). POS features were normalized by dividing by the number of tokens in the document. For the lexical features, Term Frequency-Inverse Document Frequency (TF-IDF) weights were computed. Considering the modest size of the training set, feature selection using an Extra Trees Classifier was performed and the top 300 features were selected. The analysis was performed using the majority vote classifier that weighed outputs from: (a) SVC with Linear kernel, (b) Decision Tree classifier, and (c) Multinomial Naive Bayes.

The method was tested over 10 independent train-test runs, where 10% of the data were randomly withheld, and the classifier was trained on the remaining 90% of the data set. The ensemble method was able to classify student-produced content with 93% accuracy.

4.3.4 Evaluation 3

The third evaluation employed the dataset used in the proof of concept study. The dataset comprised student-generated content by eleven students enrolled in one graduate-level research methods course.The size of the dataset was as follows: The word count of assignment 1 ranged between 1,200 and 5,700 words; Assignment 3 ranged between 1,900 and 4,500 words; Assignment 4 ranged between 800 and 3,600 words; Forum messages ranged between 250 and 450 words and between 1 and 10 messages (mode 5) per student. Documents were divided into chunks of 500 words. The feature set comprised noncon-

secutive word pairs with a window size = 10, frequency of occurrence $\geq$ 5, and counts were TF-IDF normalized. The importance of each feature was computed using the Extra Trees classifier (Geurts et al, 2006) and the feature space was reduced to the top 300 features. The analysis was performed using the majority vote classifier Raschka (2015) that weighed results from three supervised classifiers: (a) SVC with Linear kernel, (b) Logistic Regression, and (c) Multinomial Naive Bayes. All analyses were conducted using the 10-fold cross-validation method, using the 90:10 data split. The approach yielded an accuracy rate of 87%.

4.4 Summary

This chapter extended the proof of concept that employed the Multinomial Nave Bayes classifier and a bag-of-ngrams approach. The proposed approach employed spanning intervening bigrams (non-contiguous word pairs)—alone and in combination with POS tags—the majority rule voting, and a tree-based feature selection technique. These corpora are representative of the real-world scenario. To assess the relative performance of the proposed approach, it was imperative to know the level at which human instructors are able to accurately classify student writings. To this end, a study to measure the performance of instructors classifying student writings by author was conducted.

The performance of the computational method was higher than that of the human instructors. The results are depicted in Figure 4.3. The size and the type of data employed are a limitation of this study, as are the feature sets and the algorithms used.

The computational approaches discussed in this chapter provide the means for mapping learner identities with their academic work. These approaches perform user-level
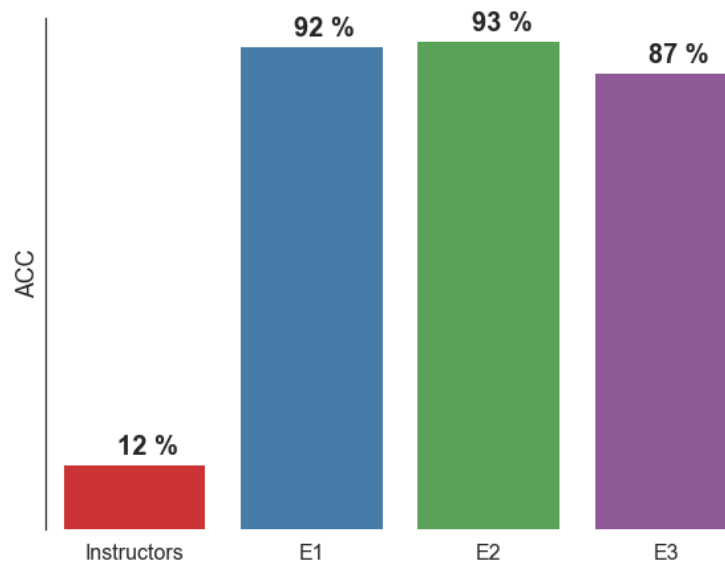
Fig. 4.3: Performance results

analysis with an aim to provide identity and authorship assurance. The available student-generated content may also undergo content-level analyses, expanding the scope of application from academic integrity only to e-assessment. Both types of analysis can be integrated into a system whose aim is to streamline the evaluation of student work and provide a variety of insights. This point will be further explicated in the following chapters.

Part III

Pragmatic Perspective

This part of the thesis comprises three chapters sharing a theme of pragmatic innovation. Chapter 5 presents the prototype application aimed to empower academic practitioners with a more efficient means for maintaining academic integrity and provides students with convenient and non-intrusive ways to complete assessments. The prototype embodies a system for a computational academic integrity approach. The system architecture is modular and allows for additional analyses to be incorporated that expand functional application from being strictly academic integrity oriented to facilitating a variety of e-assessment tasks. To demonstrate this capability, a method for thematic and structural visualization of academic content was developed and integrated to provide content-level analysis. The method is presented in Chapter 6. A procedural-based framework for integrating data-driven applications into institutional practices is presented in Chapter 7, using the academic integrity task as an example.

Chapter 5

Open Proctor: An Academic Integrity Tool

Overview

The challenge of managing integrity of assessments in an open and distance environment stems from a need to map student identities with their academic work in effective and efficient manners, while preserving privacy, ensuring minimal disruption, and minimizing impacts on accessibility and convenience. Theoretical understanding of the problem—supported by the empirical data— has dovetailed into a pragmatic approach that uses machine-learning techniques to analyze patterns in the learner-produced academic content. A cloud-based application entitled OpenProctor which analyzes patterns in the learner-produced content was developed. It provides both identity and authorship assurance in a single-tiered process and produces a simple to understand report flagging potential cases of academic misconduct. OpenProctor puts the student in the center of the learning process and does not seek to control the remote learning environment or impose restrictions on the way the students learn. In the following sections, the system architecture and functionality are discussed, and the instructor perceptions towards it are examined.

5.1 System architecture

OpenProctor has a modular architecture comprised of two components and five modules.
The two components include: (a) the data management interface and (b) the analytics en-
gine. The five modules include: (1) user management, (2) learning task management, (3)
student profile management, (4) validation of academic integrity, and (5) a messaging ser-
vice. The function of the data management interface is to manage the student-generated
data and pass it on to the analytics engine. The function of the analytics engine is to
process the data, discover and analyze patterns within, and generate the integrity report.
These functions will be discussed in the next sections. The process diagram is depicted
in Fig. 6.1.
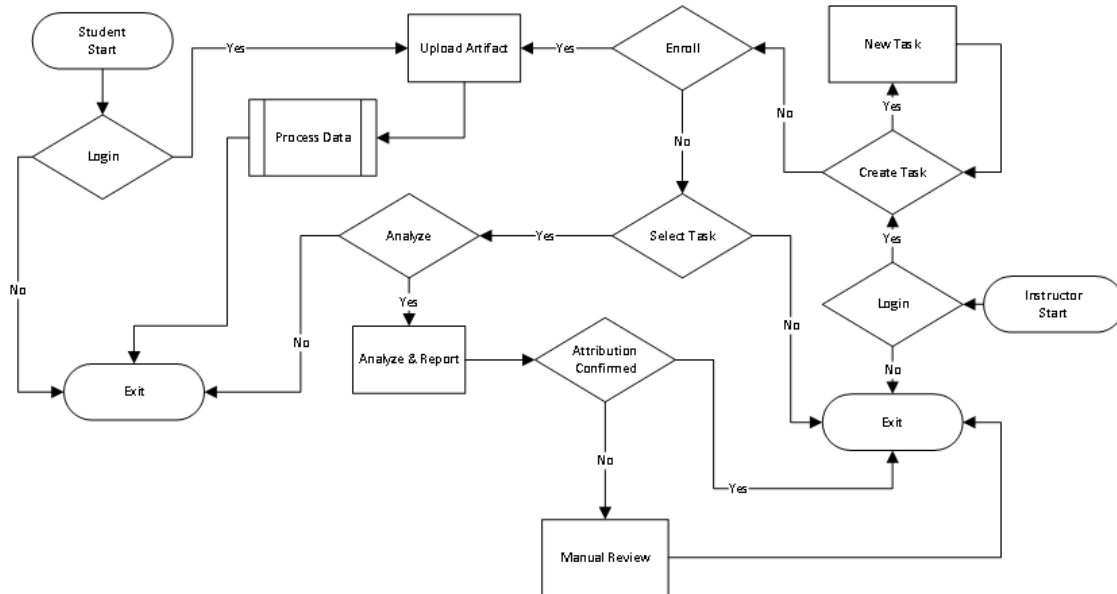
Fig. 5.1: Procedural diagram

The current version of the OpenProctor has been implemented on a LEMP stack: Linux, Ngnx Web server, MySQL database with the Laravel PHP framework powering the data management functionality, and Python machine learning and natural language libraries (Bird, 2006; Pedregosa et al, 2011) powering the data processing and analysis modules. A high-level system architecture is depicted in Fig. 6.2.
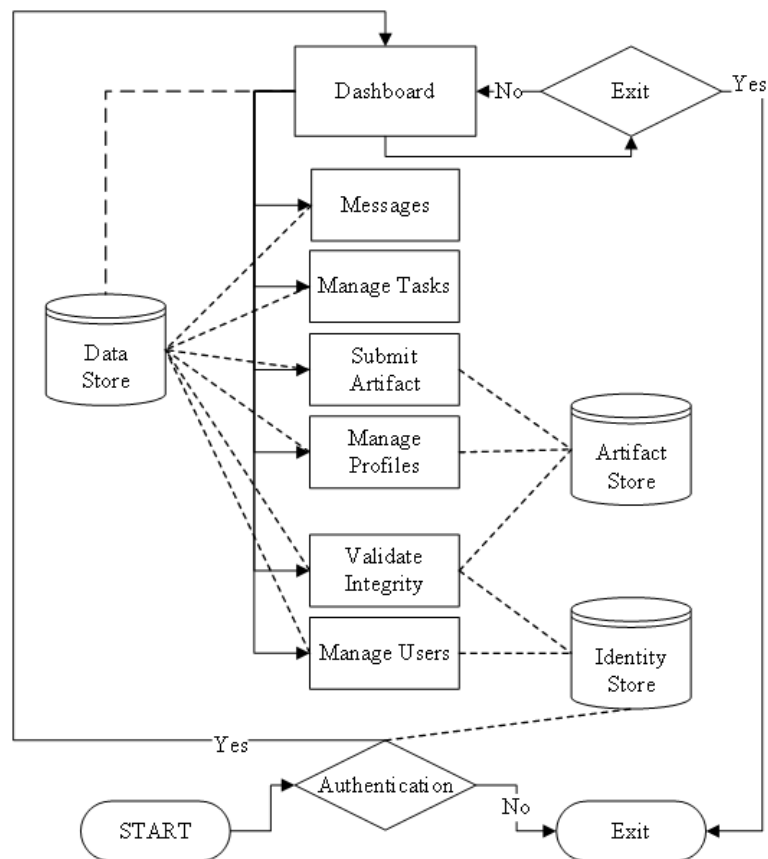
Fig. 5.2: System architecture

5.1.1 User management

There are two user types in the OpenProctor system: the standard user (the student) and the administrative user (the instructor). The instructors manage core aspects of the system such as user management, learning task creation, generating reports, among others, while the students can upload their artifacts for each of the learning tasks. User accounts can be created by the instructor: by specifying the student's name, e-mail, password, and the type of the user account; and also by the student: through the account self-registration process, which by default makes all self-registration accounts to be of the standard user type. User passwords can be reset by the instructor or through the password recovery process.

5.1.2 Learning task management

Courses are composed of one or more learning activities, some of which culminate in production of student-generated content such as the written assignments or the online group discussions. These academic artifacts constitute minable data, containing author-specific patterns that can be used to validate the identity and authorship claims. Because not all learning activities bear an equal grade weight, the ones that are a part of the assessment process, need to be identified. For example, the students might be invited to participate in the weekly online discussions throughout the course duration, but only the discussions that address specific issues are counted towards the final grade. Although the entire discussion thread is collected, only a part of it will be used in the assessment process and needs to be defined as such. The task management function provides for doing just that.

It is designed to segment the mapping of learner identities with the student-produced artifacts for each activity that is a part of the student assessment process.

Each validation task in the OpenProctor system corresponds to the assessment task in a particular course. If a course is composed of two graded assignments, there will be two tasks and two reports, one for each assessment item. The task has a name and status of being active or inactive. They can be activated, deactivated, renamed, created, or deleted at any time during the course. Uploading student assignments is contingent upon task creation by the instructor. Students can only upload content as a response to a specific task, and therefore the assessment task creation by the instructor supersedes the work submission by the student. Once the data is uploaded, students can view, delete and resubmit their work. The student-level analyses are performed on a per-task basis, where one report per assessment task is generated.

5.1.3 Student profile management

In order to validate authorship claims of the submitted artifacts, one needs to have a basis of comparison. The student profile provides just that by enabling the instructor to select the learner-generated content that subsequent student work will be validated against. Student profiles collectively constitute the training set that the supervised machine learning algorithms are trained on, and the student's identifier is the class label. Only the instructor has access to this function, and the data in the profile may be updated as frequently as the instructor deems necessary. The data is expected to be as close to the ground truth as possible, that is, having high degree of confidence that the content was produced by the student who claims authorship is necessary for maintaining validity.

5.1.4 Integrity validation

Students access the OpenProctor system and submit assignments for each specific learn-ing task. The assessment process is non-anonymous and requires an identifier such as the student's name which is a part of the user profile. Students are expected to complete and submit their own work, therefore identity and authorship claims are assumed to be of the same entity. The current configuration allows one submission per learning task.

The assignments are grouped together by the learning task and collectively form a testing set which can be validated against the data in the student profile—the training set. The instructor can perform two actions: first, to delete student submission and thus exclude it from the analysis and second, run the analysis which in turn generates a report depicting classification results and highlighting cases that require instructor attention. As students progress through the course materials and submit new content, the instructor performs the analyses for each task and issues the feedback either through the internal messaging system or any other means as necessary.

The instructor receives two types of reports: The dashboard depicted in Figure 5.3 pro-vides a general overview including the number and type of the system users, the number of active and inactive learning tasks, the size of the student profiles (training set), and the number of the submitted artifacts (testing set), the most recent files and their GeoIP lo-cation. The student-level report is based on the classification analysis that flags cases for the manual review. It identifies the cases of potential academic misconduct, something an instructor should look into further while performing the qualitative assessment (grading) of the submitted work. This functionality is not intended to relieve the instructors from being vigilant of possible cheating, but rather to provide an additional layer of checks

and balances that helps to examine the content in quantitative terms. The report is only a suggestion to scrutinize the flagged cases, and it is still at the instructor's discretion to perform the additional checks.

Fig. 5.3: Dashboard
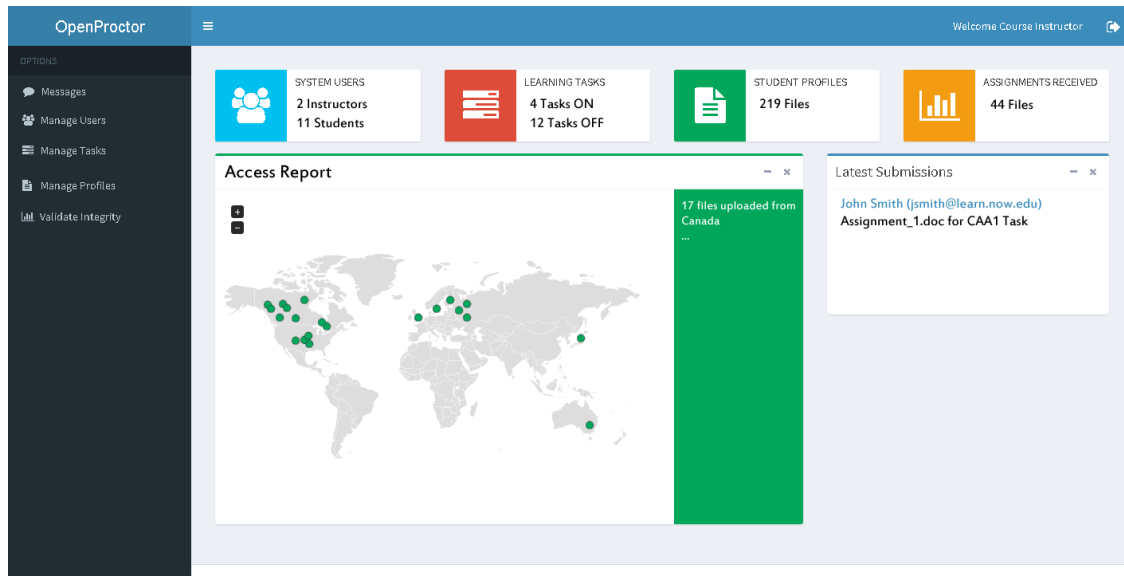
5.1.5 Messaging

The messaging component facilitates the exchange of private and public messages among the users. For example, the instructors may post updates, make announcements, or provide students with feedback. Students may communicate their concerns or provide additional details with regards to their submissions. This module can be used alongside the message board of the LMS.

5.1.6 Analytics engine

The raw data—the assignments that the students submit— are human readable and dis-seminated in a variety of file formats. These need to be made machine readable and therefore, formatted according to the input requirements of the machine-learning algo-rithm. The data undergoes preprocessing which removes the noise, tokenizes text, counts tokens, normalizes token counts, and turns them into numerical vectors. The feature vec-tors are fit with the classifier which returns a class prediction. Upon the analysis a report is generated, suggesting which artifacts the instructors should pay closer attention to when performing a qualitative review. The report comprises the following information: The number of students/assignments included in the assessment task, a list of students who work require additional review, classification matrix, name of the assessment task, and a timestamp.

The computational tasks were implemented using Scikit-learn library (Pedregosa et al, 2011) in Python language. Modular architecture enables easier scaling of the analytics engine to include new algorithms and data processing functions.

5.2 Instructor perceptions

The previous sections discussed the components of an application for maintaining iden-tity and authorship assurance in the learning environment. The adoption of innovation is often contingent upon stakeholder buy-in. It was important to examine the perceptions of instructors regarding the analytics-based approach to academic integrity. To this end, a survey was administered to the instructors, following a demonstration of the technology.

OpenProctor was showcased at the second world conference on blended learning (IABL2017). The venue provided an opportunity to connect with and discuss the issue of academic integrity with academic practitioners and administrators.

OpenProctor received its own booth at the conference where the participants visited the booth between sessions and were given a demonstration of the system. The demonstration emphasized the main differences between observational-based approaches, plagiarism detection, and behavioral-based strategies. The demonstration of the technology was performed using tablet computers. Notes were taken. Due to the busy nature of the event, administration of the survey was challenging, and much of the data collection was performed outside of the conference hours by sending follow up e-mails to the participants. Additional participants were recruited using public information on university websites and by directly approaching colleagues at the International Conference on Advanced Learning Technologies (ICALT2017). Initial attempt to conduct individual demonstrations proved to be difficult due to scheduling and logistical challenges. The solution was to record a video demonstration. A video link was provided to the participants together with a whitepaper outlining the rationale and main principles of the approach.

The inclusion criterion for participants was that they were currently teaching at the college or university level. The participant group size, the convenience sampling, and the type of demonstration methods are the limitation of the study. The survey comprised 10 questions (1 qualification question, 6 five-point Likert-scale, 2 Yes/No questions, and 1 multiple-choice question with comments.) The survey aimed to examine the instructors' perceptions of the technology and the approach to academic integrity. An excerpt of the survey instrument is presented in Appendix A.4.

There were a total of 9 completed surveys; the response rate was lower than antic-ipated. The results can be divided into two parts: the direct interaction and the survey data. One notable pattern observed among the participants at the conferences, before they were provided with information about the proposed technology and were shown the system in action, is that their views of what academic integrity technology is and does were predetermined. Many of the participants were eager to guess what the technology does before it was introduced (e.g., "you are like ... [a commercial plagiarism detec-tion tool]"). Surprisingly, the automated and remote proctoring services did not come up as one of the comparisons. Many of the participants associated academic integrity with plagiarism. This necessitated a need to explain that academic integrity is more than just plagiarism but comprises two distinct layers— identity and authorship. A thought experiment was conducted where the participants themselves were hired as ghostwriters, —which constitutes a case of academic misconduct, but where plagiarism detection tools are ineffective. At that point, the participants generally agreed that plagiarism detection tools have limitations and a more comprehensive strategy is required to deal with un-bridled creativity of students determined to cheat. Another common theme that emerged was that the participants were concerned with academic integrity and were on the lookout for the new ways in which to bolster the integrity of their courses.

The survey results are summarized in Figure 5.4. The results provide preliminary sup-port that the proposed approach received a positive response from instructors; however, they were seeking more information and an opportunity to interact with the technol-ogy. Four instructors consider the method compatible with the way they teach, and one instructor did not find it compatible. Five instructors were interested in using the tech-nology in their courses. Four instructors perceived that the technology promotes account-

ability, and one participant disagreed. Four instructors perceived that it promotes convenience. The strategies that the surveyed instructors employed in the past year to maintain integrity in their courses included plagiarism detection tools (Turnitin, SafeAssign), instructional design, manual search engine queries, and proctored exams that require "students leave belongings outside testing room." Six instructors believe that it is possible to detect academic misconduct without the use of technology. The majority of instructors reported not employing analytics in their courses/institutions; however, they perceived no difficulties in being able to explain to others the advantages of using analytics for providing academic integrity. Only two instructors reported having control over strategies and tools they can use to provide academic integrity in their courses.
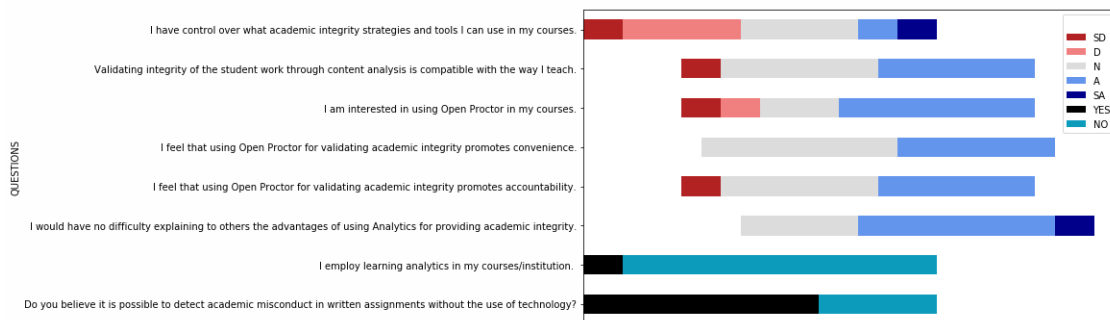

Fig. 5.4: Summary of instructor perceptions

5.3 Summary

This chapter presented a prototype cloud-based academic integrity system entitled Open-Proctor. Its student-centered approach to academic integrity does not attempt to control the remote learning environment, but aims to provide flexibility and convenience of any-

time and anyplace learning. In contrast to the traditional academic integrity strategies that aim to restrict usability and monitor user actions, OpenProctor maps learner identities to the academic work through analysis of patterns in the learner-produced content. Assessments should not be stressful or logistically troublesome for learners or instructors, and the proposed approach strives to deliver just that. Instructors' perceptions regarding OpenProctor were examined and the results suggest a positive response. The next chapter attempts to expand the scope of analysis performed by OpenProctor and add content-level analysis to simplify the task of reviewing academic content. This would transform the system from academic integrity only tool to a modular e-assessment platform.

Chapter 6

ThemeTrack: A Method for Thematic and Structural Visualization of Academic Content

Overview

The focus of the earlier chapters was on the user-level analysis that enables verification of authorship claims and mapping of the student identity to content. The aim of this chapter is to propose a novel content analysis method, which when embedded into the Open-Proctor system will expand its functional scope. The method is termed ThemeTrack and allows users to create a visual representation of the textual data, obtain a succinct summary of the information it carries, and draw comparisons between them. It also allows the user to get a snapshot of the document without reading it in its entirety and has several pragmatic implications for the process of learning and teaching. First, it enables instructors to create visual content summaries. Second, the method may serve as a visual aid for language teaching. Third, the method can be used to make inferences about the document composition.

Much of the assessments are based on the quality of written content such as essays, portfolios, and discussion forum participation. Content analysis, such as survey of literature and review of the student assignments is often a time consuming and cognitively-

demanding task for students, researchers, and instructors alike. The proposed framework aims to make the review of written content more efficient.

Academic writing is structured. In general, the problem that the researcher is trying to address is stated at the beginning of the paper, followed by what is already known— the related work. The solution is introduced approximately half way into the paper and compared to the related work towards the end–the discussion section. This suggests that different themes are introduced at different times and some themes may overlap.

The process of writing is sequential. A chunk of text does not appear spontaneously, but is formed gradually by connecting words into coherent and rule-guided structures. Writing is like threading beads on a string and if placed on a time line, each word represents a point in time, or using an earlier metaphor, a bead on a string. Since the direction of writing is known, it is possible to identify positions of words, themes, notions and concepts relative to each other or parts of the document. Some words, notions and concepts are auxiliary, whereas some are key to the argument and are carried throughout the text. By identifying tokens (that represent themes) and their positions in text, it is possible to thematically separate textual data into sections, identify main themes and delineate related concepts.

6.1 Background

Visual representation of textual data can be classified into three categories: quantitative, contextual and semantic. The quantitative approach represents text as term counts. Terms could be defined as individual words, or co-locations such as bigrams and trigrams. A common textual visualization technique that uses term counts is the word-cloud (Figure

6.1). The idea behind the word-cloud is that the term frequency determines the term visual properties such as the font size or the color. Function words are often removed as they are frequent and noisy.

Fig. 6.1: Word Cloud visualization of this article

This approach has been applied in the academic setting and integrated with informal assessment (Kitchens, 2014), providing the means to visualize and compare students' understanding of course content. One may critique the utility of word-clouds on the basis of their inherent limitation to deliver only a shallow representation of the textual data.

Textual data can also be plotted as a time series graph to depict the relationship between the time and token frequency of occurrence. For example, Google Ngram Viewer[1], is a visual information retrieval interface to a corpus of over five million digitized books (Michel et al, 2011). It provides a visual representation of the relative frequencies of word collocations in literary works. It is a useful tool for conducting research on social trends

(Michel et al, 2011) as well as linguistic research (Lin et al, 2012). Figure 6.2 depicts the frequency of use of terms: "radio" and "internet" over a period from 1900 to 2000. According to the graph, the term "radio" emerged in literary texts in the early 1900s, peaked in the 1940s and went into decline thereafter. The term "internet" became a subject of growing attention in the early 1990s and the term use has continued to rapidly increase during the next decade.

Fig. 6.2: Google Ngram Viewer

Terms can also be visualized in a context, relative to other terms. In contrast to the word-clouds, word-trees[2] provide the means to examine the term relationships (Wattenberg and Viégas, 2008). Figure 6.3 depicts words and phrases that follow a root term. This approach has been found useful in literary analysis. Similar to word-trees, tag-clouds depict relationships between the terms (Vuillemot et al, 2009). There are different variations

---

[1] Created online at https://books.google.com/ngrams

of the tree structures offering different functionality. For example, double-trees visualize terms in context as two-sided trees, and are used in linguistic research (Culy and Lyding, 2010). They can include the term frequency information for both the words in context and the branching factor. This method of visualization is employed in corpus linguistics allowing to efficiently delineate differences between texts (Magnusson and Vanharanta, 2003).

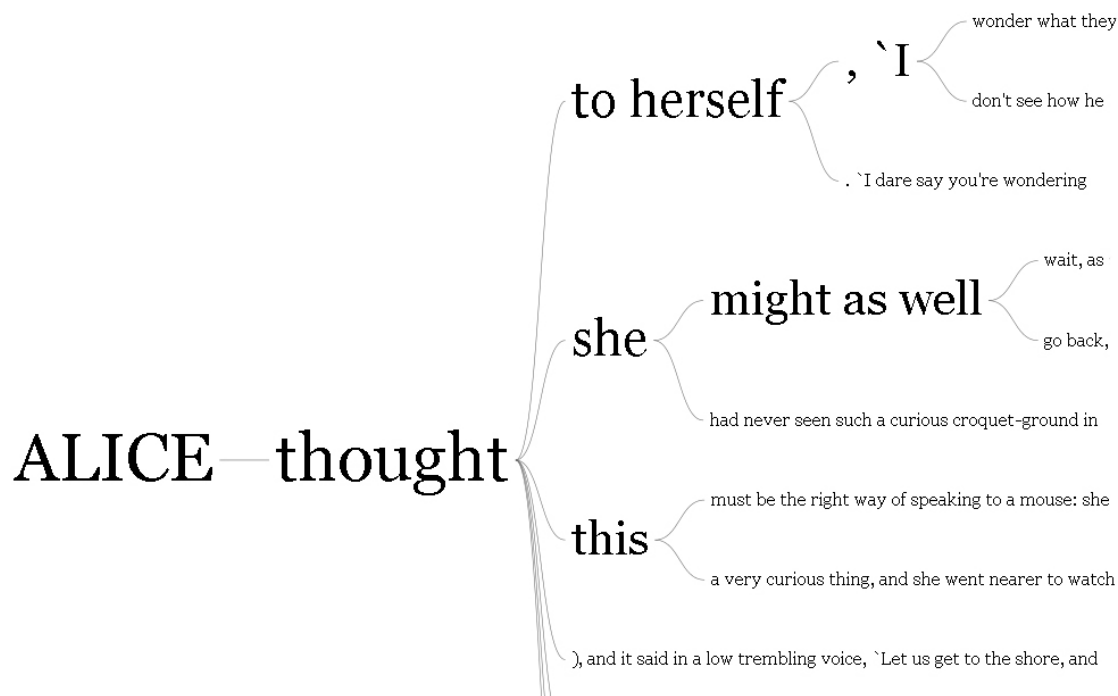Fig. 6.3: Word Tree visualization of Alice's Adventures in Wonderland

The third type of visualization approach is based on the semantic representation of text. Unlike the previously discussed approaches that organize textual data based on the frequency of occurrence or collocation, semantic-based visualization organizes terms and

---

[2] Created online at https://www.jasondavies.com/wordtree

their relationships based on their meaning. For example, Directed-graphs (Rusu et al, 2009) depict semantic structures by extracting subject –verb –object triplets from each sentence and attaching WordNet (Miller, 1995) synsets (related terms). This is attained through POS parsing and extracting named entities. DocuBurst[3](Collins et al, 2009) is a radial graph that depicts an IS-A relationship of a term by utilizing the noun-verb hierarchies of WordNet and term frequencies. A DocuBurst visualization of Leo Tolstoy's War and Peace with war at the root is depicted in Figure 6.4.
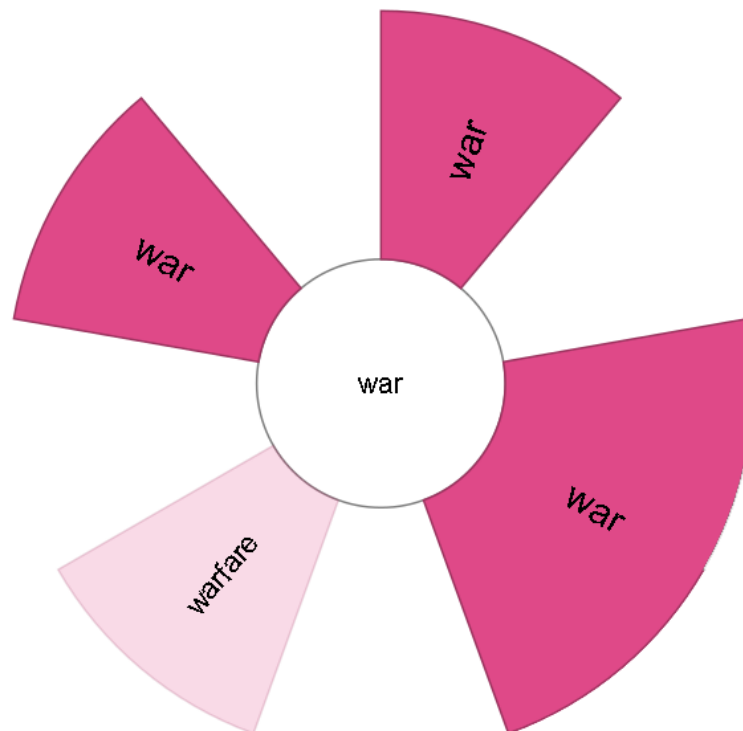

Fig. 6.4: DocuBurst visualization of War and Peace

---

[3] Created online at http://vialab.science.uoit.ca/docuburst

6.2 Proposed method

This section presents a method for thematic and structural visualization of textual data titled ThemeTrack which maps cumulative token counts to their relative position in text. In contrast to the existing approaches, ThemeTrack depicts the relationships among user defined terms: their emergence, co-occurrence, and decline. The terms may be expressed as lexical features (e.g. word bigrams), they may be expressed as syntactic features (e.g. POS), and also as semantic features (e.g. named entities, sentiments). ThemeTrack visualization can be applied on a single document or a corpus of documents by the one or more authors; and also applied to a class of documents sharing some criteria (e.g. author, subject, genre, etc.). The method allows the user to see how a text is written. It can be considered a visual disassembler that depicts (researcher defined) components of a text; it depicts how the information is organized and presented. The method allows the user to see, in quantitative terms, how one property of a document is related to another. For example several book volumes of one author could be analyzed to identify recurring themes or the use of literary devices within each book and across the volumes. It may also be adopted to analyze other textual data such as music scores. It attempts to answer the questions: what is in the text, and how is it all put together? This will become more obvious from the examples in the following paragraphs.

The method is comprised of six steps: (a) pre-process text, (b) extract information, (c) count tokens, (d) identify token positions, (e) create pairwise mappings between cumulative token occurrence and their positions, and (f) plot a correlation between the token position and cumulative token count. The output is the two-dimensional x-y graph showing cumulative term frequency on the Y axis and the term position in text on the X axis.

The token position is a distance between two events. It can be defined as a temporal dimension (to see which events occur at the same time or over time) or a spatial dimension (to see how far an event is from any part of the document.) The argument follows a sequential order and has a clearly defined beginning and an end. Academic texts are comprised of multiple themes; their relationships can be established at any point in text by measuring co-occurrence.

Figure 6.5 provides an illustrative example of ThemeTrack visualization. In this example, there are three distinct themes. The main theme is present throughout the document (it starts at the position 0 and continues until the very end of the document) and has been repeated 28 times. Theme # 2 is introduced later in the document, and repeated 6 times and does not reoccur later. Theme # 2 is related to the main theme because they co-occur together. Theme # 3 is introduced towards the end. It is not discussed in context of Theme # 2, because there is no overlap, but is related to the main theme. Its frequency of occurrence is higher than that of Theme # 2, so is the time spent discussing it, and therefore it is more prominent than Theme # 2.

Plotting the cumulative frequency of term occurrence and their position produces a two-dimensional x-y graph. The themes are sorted by frequency and the N most frequent themes are plotted. The number of themes is user-defined. For every unique and new token occurrence, the token count increases by 1. The token position (X axis) can be expressed in terms of text length, term count, or as its percentage. Depending on the scale, it can answer questions such as: After how many words a term is repeated, or at what point certain themes converge.
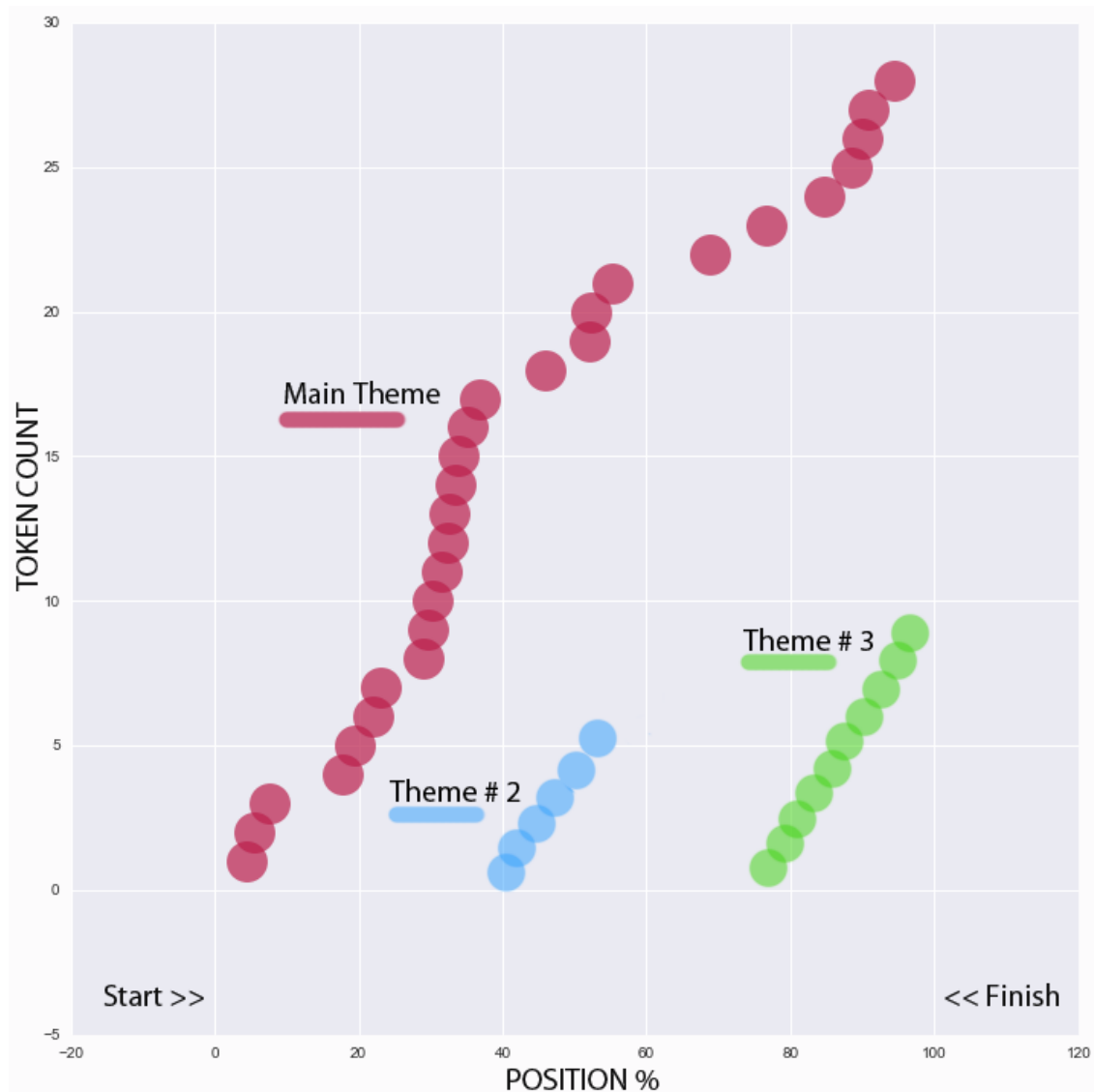
Fig. 6.5: ThemeTrack visualization

6.2.1 Data processing

The raw content comes in a variety of formats and uses various templates. For example, much of the journal articles are distributed in the portable document format (PDF), so are the theses and dissertations, whereas much of the student assignments are distributed

in editable formats such as Word documents. The paper layout of journal articles and academic courses vary in the templates they use as well as the style of the bibliographic references. These differences need to be taken into account when processing the raw text, because repetitive headers, footers and citations will contribute to the noise. Once the text is free from noise, it undergoes the information extraction step whose aim it to delineate the main themes and their positions in text. The natural language processing (NLP) techniques provide just that. This can be performed in a variety of ways: ngram based methods, shallow and deep NLP techniques. For example, tokens may be comprised of single words, ngrams (contiguous words or spanning intervening words), POS tags or syntactic ngrams, semantic clusters, etc. In spite of the variations in the information extraction protocol, the underlying concept of the proposed approach is to depict the token lifecycle in relation to other tokens.

The following sections discuss the analyses that employ ngram based methods.

6.2.2 Algorithm

The proposed method bears similarity to the frequency based approaches in that the themes are quantified and the magnitude is depicted on a graph. It also bears similarity to the contextual approaches in that the graph depicts the high-level relationships between the themes. It paints a picture of what themes are co-occurring at any point in the text. The proposed method creates a view into the token lifecycle—emergence, reoccurrence and decline.

The method is algorithmically expressed in Figure 6.6.

Fig. 6.6 ThemeTrack textual data visualization

1: Open document

2: Parse document into plain text

3: Remove noise

4: Select information extraction method

5: Tokenize text

6: Create frequency-position matrix

7: Sort by most frequent token

8: **For** each token plot cumulative token count and its position **end For**

9: Close document

6.3 Method evaluation

An exploratory analysis was conducted using a corpus of the real-world academic texts composed of 20 student thesis and 20 published journal articles. The texts were between 3,000 and 85,000 words. The five most frequent terms were plotted. The thesis dataset was comprised of 10 doctoral-level theses obtained from The Digital Archive of Research Theses of the Open University UK [4] and 10 theses at the master's level, obtained from The Digital Theses library of Athabasca University[5]. The second dataset was comprised of 20 journal articles obtained from the IEEE Xplore Digital Library [6] covering a variety of computer science topics and from The International Review of Research in Open and Distributed Learning[7] covering research in distance education.

Computational and graphing tasks were performed using the Python programming language: using standard libraries for parsing documents into plaintext, regular expressions

---

[4] http://www.open.ac.uk/library/library-resources/theses-dissertations

[5] https://dt.athabascau.ca

[6] http://ieeexplore.ieee.org

[7] http://www.irrodl.org

for filtering out the noise, and the NLTK library (Bird, 2006) for the natural language processing tasks were employed.

The aim of the experiments was twofold. First, to assess the viability of the proposed visualization technique using a dataset of the real-word academic texts. Second, to compare visual representation of information extracted using ngrams to represent themes, and syntactic parsing technique utilizing verb-noun pairs to represent actions. The identified themes and actions were compared against the titles and keywords (for journal articles), which were removed during the pre processing, as were the bibliographies.

6.4 Results

Both methods yielded visually similar results, although word ngrams captured more meaningful information in the sense that actions (operationally defined as verb-noun pairs) were capturing a lot of abstract information (e.g. giving rise), and therefore insufficient by themselves to yield any concrete inferences about the nature of the text.

The extracted themes were in-line with the keywords and titles. Much of the journal articles exhibited a consistent pattern with one dominant theme carried throughout the document, and supporting themes that emerged and declined, while much of the theses had multiple related themes carried throughout the manuscript. In other words, journal articles exhibited a more sporadic flow of ideas. A sample visualization of a doctoral thesis titled "How does the use of mobile phones by 16-24 year old socially excluded women affect their capabilities?" (Faith, 2016) is depicted in Figure 6.7. A sample visualization of a journal article entitled "Effective pattern discovery for text mining" (Zhong et al, 2012) is depicted in Figure 6.8.
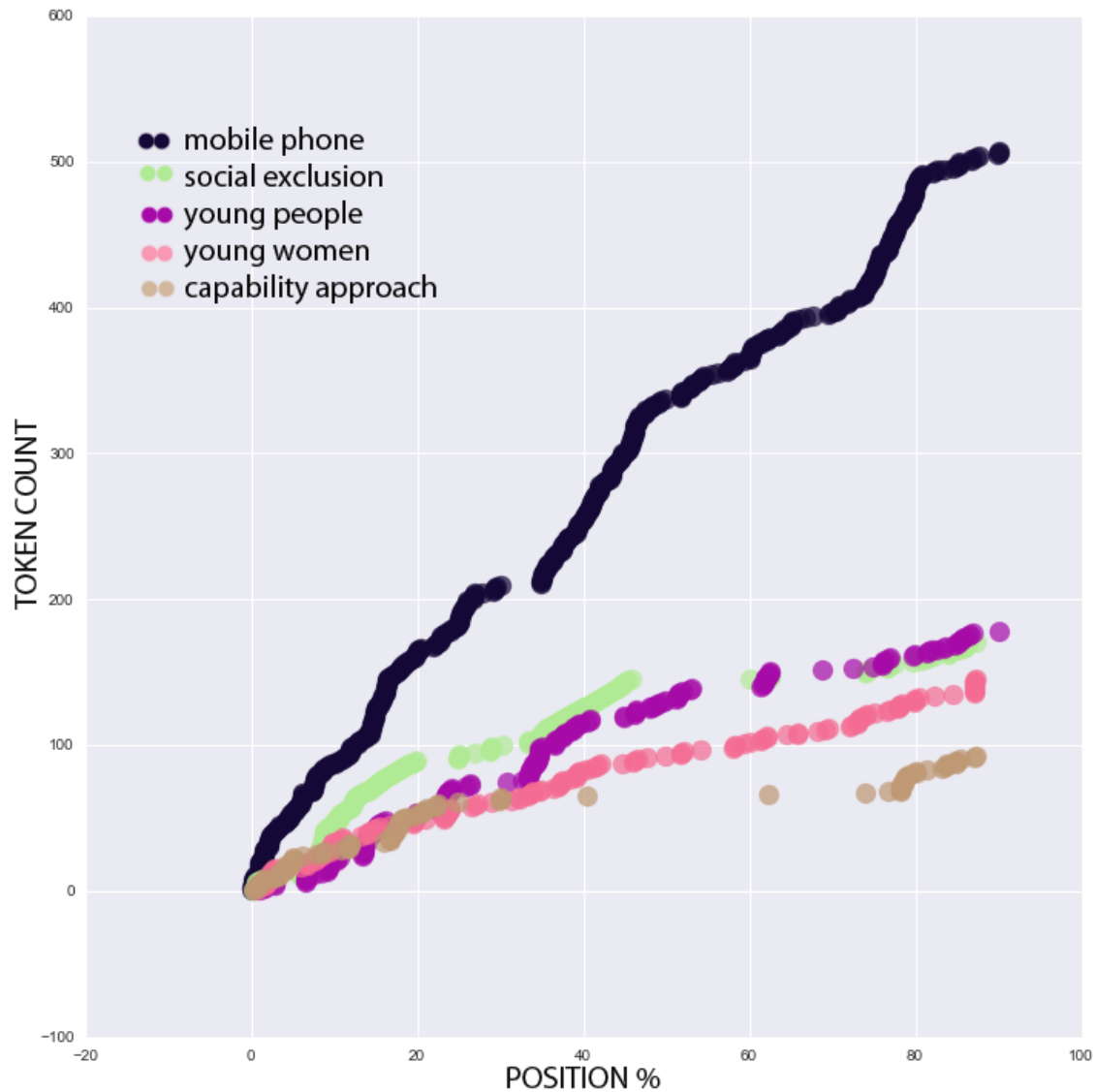
Fig. 6.7: ThemeTrack visualization of a sample doctoral thesis

These visualizations provide a succinct summary of how content is structured. In the journal article, the terms "text mining", "discovered patterns", and "pattern mining" are semantically similar and occur throughout the document, whereas the term "closed patterns" is introduced at the beginning (around the literature review section) and later rein-

Fig. 6.8: ThemeTrack visualization of a sample journal article

troduced at the end of the document (around the discussion section). The authors are

making a connection between what is new, and what is already known. The notion of

"positive documents" is central to this article, it starts after the introduction and merges

with the theme of "pattern mining" and "text mining". The size of the journal article is smaller than the thesis, therefore the data points look sparse.

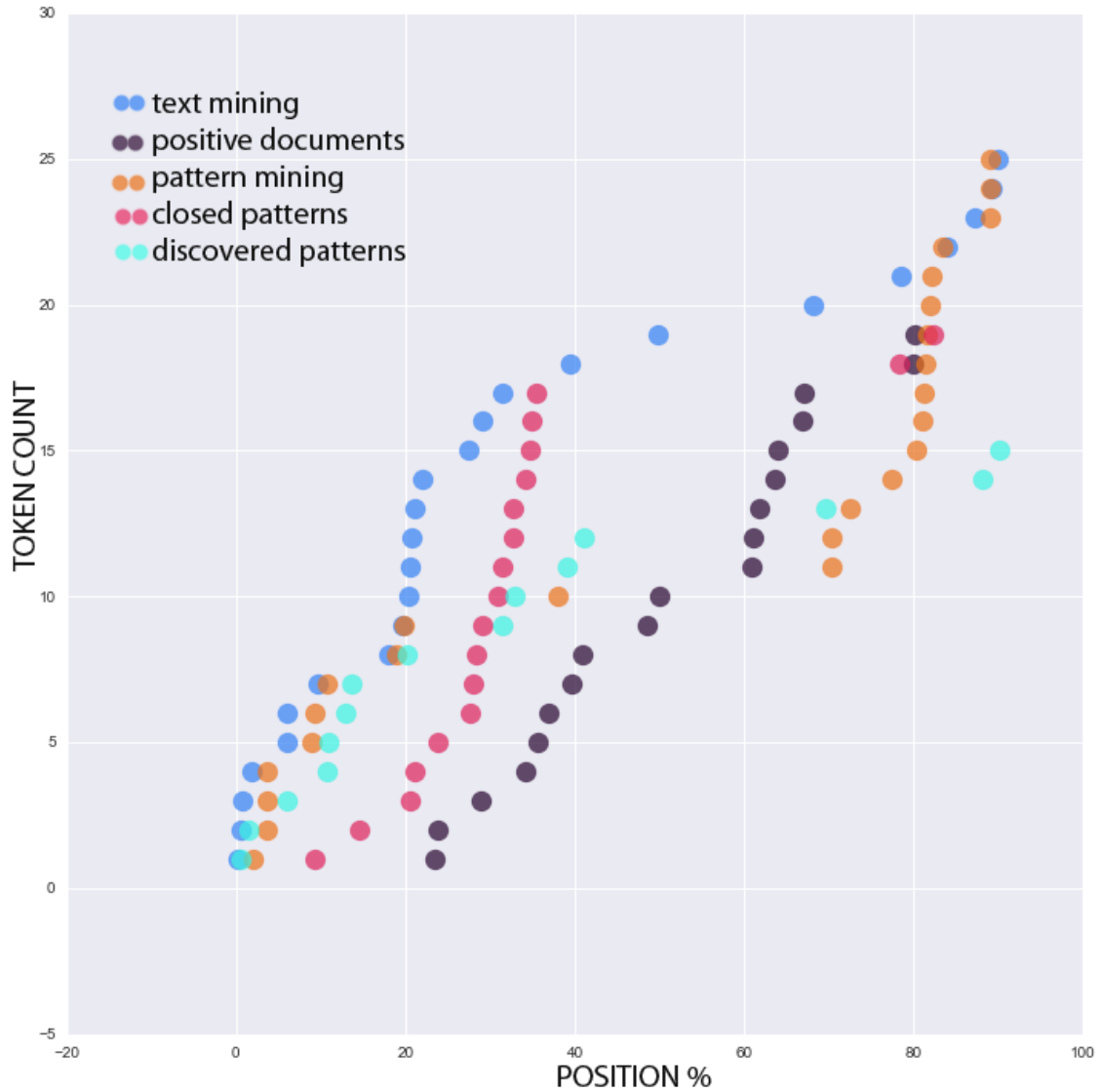In contrast to the journal article, the sample thesis has a different organizational structure. Figure 6.7 shows that all five themes continue throughout the document, with "mobile phone" being the most frequent term. The term "social exclusion" is the second most frequent term at the beginning, and at around half-way through the document, the term "young people" takes its place. The terms "young people" and "social exclusion" show an overlap that is they are both used with similar frequency. The term "capability approach" is introduced in the beginning, there are a few instances where it is referenced in the middle of the document, and then reintroduced towards the discussion and the conclusion.

6.5 OpenProctor integration

Themetrack algorithm was integrated into the OpenProctor system to provide content-level analysis. Figure 6.9 depicts the updated architecture that includes content-level analysis and Figure 6.10 presents the dashboard and content analysis module.

Analysis of the student-generated content produces an aggregate report for all works submitted by each student. This enables the instructor to compare and examine the progression of student learning. By creating a visual map, the assessment process is simplified in that the instructor is provided with a snapshot of the paper structure and a list of main themes. The modular architecture of the system allows additional mature and novel algorithms to be added to the analysis pipeline providing the instructors with variety of insights.
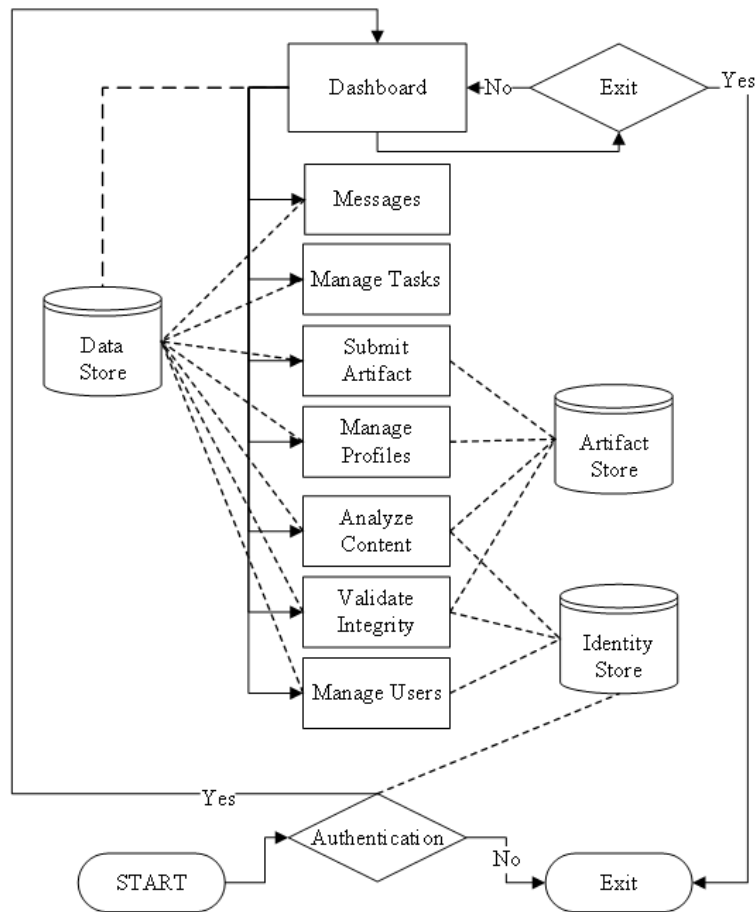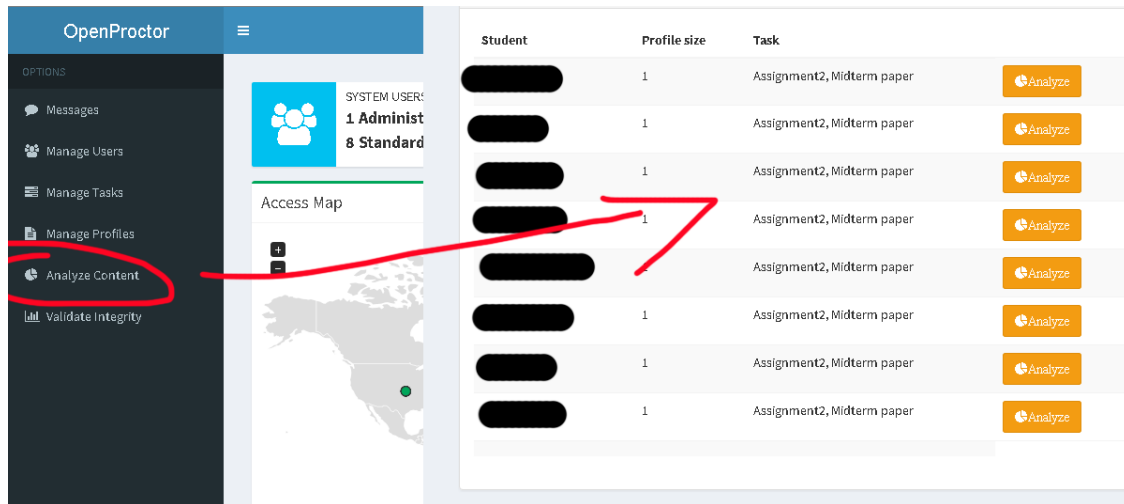
Fig. 6.9: OpenProctor architecture

6.6 Summary

The focus of this chapter was on the content-level analysis. A method for thematic and structural visualization of academic content, titled ThemeTrack, was introduced. The method is aimed to simplify the review of student-produced content and create a visual representation of the themes and structure within a text. The integration of the content-level analysis into the OpenProctor system expanded its functional scope from being strictly an academic integrity tool to an e-assessment platform. The updated OpenProctor

Fig. 6.10: OpenProctor integration

system enables instructors to create visual content summaries and use visualizations as teaching aids.

The previous two chapters discussed the development of data-driven solutions to academic problems. The following chapter presents a framework for the integration of these approaches and applications into institutional practices.

Chapter 7

Creating a Roadmap for Implementation

Overview

The previous chapters discussed data-driven approaches that aim to enhance the integrity and convenience of the student assessment process. However, availability of technology does not entail its successful implementation or acceptance. The literature is abundant with examples of great proposals that did not reach a production phase, because they did not include a roadmap for implementation. This chapter proposes a framework for integrating data-driven approaches into an overarching domain of academic and institutional practices. The main challenge for creating such a framework stems from the context-specific nature of the data analysis, which requires it to be able to address any environmental constraints that may arise. The strategy that has worked well at one institution may not always translate well to successfully capturing and measuring data in another. The proposed framework shares many of the considerations of earlier proposals (Campbell et al, 2007; Chatti et al, 2012; Clow, 2012; Greller and Drachsler, 2012; Khalil and Ebner, 2015) and encompasses a set of procedural steps for creating and implementing analytics tasks. Analytics is viewed as a sum of parts that may change with environmen-

tal shifts. It is also considered a needs-based measure, where each task has an owner and is designed to address a specific need, which contrasts with the one-size-fits-all approach (Larose, 2005; Gašević et al, 2016).

7.1 Background

The two main approaches for constructing knowledge from data are analytics and data mining. These, put in the academic context, become Educational Data Mining (EDM) and Learning Analytics (LA). The International Educational Data Mining Society defines EDM as: "an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in" (Siemens, 2012). The Society for Learning Analytics Research defines Learning Analytics as: "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (Siemens, 2012). A wider definition that encompasses computation has been proposed, and defines analytics as the "use of data, statistical analysis, and explanatory and predictive models to gain insight and act on complex issues" (Brooks and Thayer, 2016). Considering the broad scope of analytics and its applications in the academic context, in this chapter analytics is defined as the process of uncovering insights about the state of learning and administrative affairs by applying computational techniques on the available data and acting upon them. Its scope depends on the particular institutional context and may range from student success to financial liability to technology management to accreditation deficiencies. The analyses can be conducted at the institutional level (e.g.,

procurement practices), at the course level (e.g., quality of the learning materials), and also at the individual level (e.g., learning style).

Both analytics and data mining draw upon computer science to provide the necessary theoretical underpinnings and practical methods for working with data. The dichotomy between LA and EDM is not clear cut because in both approaches, processing, computation, and reporting functions play key roles. In much of the literature and popular media the terms analytics, mining, machine learning, and artificial intelligence are used interchangeably to describe different elements of the data analysis process, which may lead to confusion (Gudivada et al, 2016). Some posit that the EDM is a part of LA that deals with computational aspects of the analysis, noting that learning analytics takes advantage of educational data mining for the retrieval of the information (Greller and Drachsler, 2012). Siemens (2012) described LA in terms of the purpose of the analysis (e.g., determine a sentiment, analyze discourse, and predict learner success rate), and EDM in terms of analysis tasks (e.g., classification, clustering, Bayesian modeling, and visualization). However, both the purpose and means of carrying out data analyses go hand in hand, for example sentiment analysis—which falls under the umbrella of LA— is often posed as a classification task (EDM process), which by itself may be executed using Bayesian algorithms (EDM process), and its results visualized (EDM process).

Another way to delineate LA and EDM is to examine the role of hypothesis: the former is hypothesis-driven, whereas the latter explores data without a preconceived hypothesis (Baepler and Murdoch, 2010). Data mining has also been described in terms of the analysis types that provide: description, estimation, prediction, classification, clustering, and association of features in data (Larose, 2005). A literature review by Dutt et al (2017) of 166 articles on the use of clustering techniques of educational data suggests that much of

the onus is placed on the user to interpret why the clusters were formed in one way or another, which may lead to variance in interpretation. Analytics, on the other hand, has been described in terms of actionable decision making that provides: diagnostic, descriptive, predictive, and prescriptive guidance (Herschel et al, 2015).

Analytics employed by the academic institutions can be organized into two categories: "learning analytics" — pertaining to learning and teaching and "institutional analytics"— pertaining to organizational processes, and business practices (Brooks and Thayer, 2016). The latter aims at tracking operating efficiency, while the former aims to improve learning experience and also to reduce student attrition. Institutional analytics can be further divided into subcategories that focus on a specific institutional issue. For example, Educause's report on the use of data analysis in the academic environment entitled "The Analytics Landscape in Higher Education" organizes analytics into five areas pertaining to learning, business, student management, faculty performance, and degree completion (Brooks and Thayer, 2016). Table 7.1 depicts the main focus areas of analytics and its applications. The report suggests that much of the focus is directed towards operational and business processes, whereas the scope of learning analytics is limited to the tracking of learning outcomes and assessment. In spite of a growing number of proposals to use data in resolution of academic-related issues, learning analytics appears to be underutilized.

7.1.1 Applications and frameworks

With the advancements in information retrieval and computing techniques, data-driven solutions to academic issues started to draw more interest. New, intuitive tools have emerged and attracted researchers from outside of the traditionally computational disci-

Table 7.1: Use of analytics

Focus area	Type of analytics
Finance and budgeting	Business
Central IT	Business
Progress of institutional strategic plan	Business
Human resources	Business
Library	Business
Facilities	Business
Procurement	Business
Enrollment management, admissions, and recruiting	Student management
Undergraduate student progress	Student management
Student degree planning	Student management
Instructional management	Student management
Student learning (learning outcomes)	Learning
Student learning (assessment and feedback)	Learning
Other student objectives	Learning
Faculty teaching performance	Faculty performance
Faculty promotion and tenure	Faculty performance
Faculty research performance	Faculty performance
Time to complete a degree	Degree completion
Cost to complete a degree	Degree completion

plines who were eager to start exploring data generated by the academic institutions. The problem of student retention was examined in a study by Elbadrawy et al (2016) through identification of at-risk students, and predicting their performance. A study by Black et al (2008) proposed to apply analytics to promote connectedness and classroom community. A relationship between student perceptions of course community and the frequency of events in the LMS activity logs was examined using statistical methods. Analytics has also been aimed at supporting student wellbeing. One study tackled the issue of cyber-bullying, much of which occurs outside of the classroom (Nitta et al, 2013). Traditionally,

the monitoring activities are performed manually by members of the parent-teacher association, which makes detection difficult and inefficient. The dataset contains real-world data from bulletin boards. The experiments were conducted using statistical techniques and their results exceeded that of the baseline. A more elaborate approach to a school-wide bullying intervention leveraging machine learning techniques in combination with hardware such as mobile devices, heart rate monitors, and video cameras was proposed by Brahnam et al (2015). The notion of providing student academic support through learning analytics has been discussed by Joorabchi et al (2016). They applied text mining techniques in order to identify subject areas which learners find the most challenging.

Analytics can be considered a tool for measuring the quality of the academic processes, an instrument for monitoring the health of the organization, and the means for promoting an institutional agenda through automation of the decision-making process. A number of frameworks support the implementation of analytics. A framework proposed by Campbell et al (2007) portrays academic analytics "as an engine to make decisions or guide actions." It presupposes having a set of objectives that need to be fulfilled. The five steps of their model include: capture data, produce a report, predict trends, act on predictions, and refine the analysis. The framework also stresses the need for a stakeholder assessment and a division of role responsibilities, the need for appropriateness of the interventions, the need for a quality-control process to improve the outcomes, and the need for understanding challenges and risks. Clow (2012) proposed to ground analytics in a learning theory, putting more weight on the outcomes and effectiveness of interventions. A framework for implementing learning analytics proposed by Greller and Drachsler (2012) is aimed to provide quality assurance, curriculum development, and improve teacher effectiveness and efficiency. The framework comprises six dimensions:

stakeholders, objective, data, instruments, external limitations, and internal limitations. This contribution is important in that it emphasizes the limitations and stresses the need for expertise to successfully operate analytics applications; this point of view has been stressed elsewhere (Larose, 2005; Gašević et al, 2016). It also stresses that stakeholders vary in their information needs, which entails a customized approach to analytics. Another learning analytics framework proposed by Chatti et al (2012) comprises four dimensions: data and environment, stakeholders, objectives, and methods. It also highlights the variance of stakeholder requirements and emphasizes the goal-oriented nature of the analytics processes. A framework proposed by Khalil and Ebner (2015) comprises four dimensions that include: learning environment, big data, analytics, and actions; again, this emphasizes various stakeholder interests and goal-driven interventions. It also outlines eight constraints that influence the design of the analytics and include: privacy, access, transparency, policy, security, accuracy, restrictions, and ownership.

These frameworks share the view that, firstly, analytics cannot be a one size-fits-all solution as different stakeholders have different requirements, and secondly, that the analyses are oriented towards some objective. The frameworks can be broken down into four questions: what is being analyzed, why is it being analyzed, how is it being analyzed, and who is involved? Some scholars have warned about the perils of using the cookie-cutter approach to analytics and stressed that the use of analytics should be carefully planned and executed to avoid costly mistakes arising from methodologically flawed analyses (Larose, 2005). The availability of ready-out-of-the-box data analytics solutions may seem to be a viable option due to intuitive design and low cost, but making accurate predictions and generalizations requires knowledge of the computational methods and the context in which the institution operates. Even the results of descriptive analysis may lead

different stakeholders to different conclusions. What follows is that the success of analytics implementation hinges upon the stakeholder's' ability to accurately interpret and act upon the analysis results (Greller and Drachsler, 2012). On the one hand, opening the possibility of customization of the analysis parameters can expand the scope of analyses; on the other hand, this feature may lead to faulty assumptions and potential hazards. One way to address this problem is to provide multi-level access to analytics, for example, through the use of dashboards (West, 2012) where access privileges are commensurate with a stakeholder's level of expertise or organizational role.

The literature on data-driven approaches and solutions to the problems in education is abundant. However, much of it is theoretical and lacking the pragmatic dimension. The next section attempts to bridge this gap by presenting a procedural framework.

7.2 Proposed framework

This study is motivated by the gap in the literature on pragmatic methods for integrating analytics into an educational context. The aim of this chapter is to bridge this gap by presenting a procedural framework for developing and implementing analytics tasks. The proposed framework shares many of the considerations with the earlier proposals (Campbell et al, 2007; Chatti et al, 2012; Clow, 2012; Greller and Drachsler, 2012; Khalil and Ebner, 2015) and encompasses a set of processes for creating and implementing analytics tasks. Analytics is viewed as a sum of parts that may change with environmental shifts. It is also considered a needs-based measure, where each analytics task has an owner and is designed to address a specific need, which contrasts with the one-size-fits-all approach (Larose, 2005; Gašević et al, 2016).

The proposal is grounded in the assumption that analytics serves the purpose of gaining insights about the status of academic and administrative activities which lead to actions. Analytics is a set of procedural steps designed to meet an institutional goal. The notion of actionable knowledge is of paramount importance because gaining insights and monitoring trends without taking actions lacks purpose and efficiency. Although the issue tracking and trending has been argued to promote accountability (West, 2012), it cannot be the means in itself. Analytics provides the necessary means for informed decision making to control or improve one or more aspects of the learning, administrative, or business processes. It is further predicated on the assumption that the performance indicators, threshold levels, and responses are known. It would be difficult to act and achieve a desired result without knowing what the outcome should be.

The advantage of the proposed framework over the existing ones is fourfold: First, it emphasizes actionable metrics. The purpose of analytics is to take action and integrate the knowledge-based decisions back into the context. Second, the framework is flexible enough to accommodate any computational approach or data management strategy. Third, the framework is needs-based and each analytics task has a clear objective and an owner. Fourth, the framework is modular, which allows for expansion and adaptation to a variety of environments. The framework is depicted in Figure 7.1 and comprises four steps and three layers.

The first layer is the institutional context, which stipulates the objectives and financial, legal, social, and ethical constraints. The need to fulfill the objectives comes from the environment in which the stakeholders operate, so do the limitations on the types of interventions that arise from the analysis of the state of the environment. The second layer comprises the procedural steps for monitoring and acting upon the changes in the state of
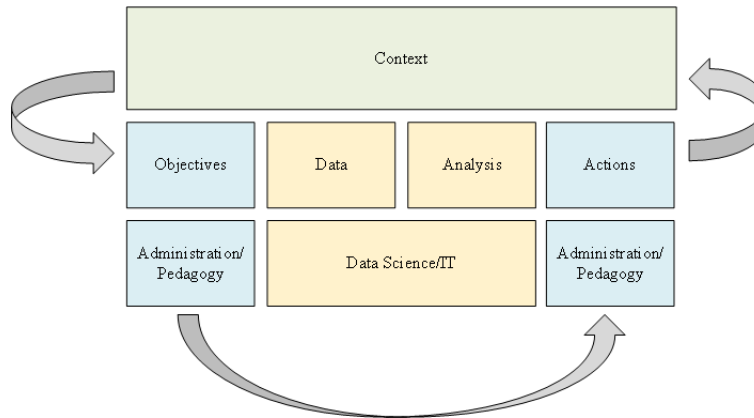
Fig. 7.1: Conceptual framework

the context. The information, both expected and actual, are drawn from the environment, analyzed, and turned into actions when necessary, which in turn are incorporated back into the environment. The third layer denotes the functional roles. Data and analyses are the prerogative of the data scientists—a group of stakeholders responsible for implementation and support of analytics, whereas the management of objectives and execution of interventions fall on the shoulders of the faculty and staff. The needs analysis, formulation of objectives, and actions are the administrative functions, whereas the functions of managing and processing information that supports the actions are the products of the data science and information technology. This suggests a need for a strong relationship between the stakeholders as well as a mutual understanding of the contextual peculiarities and overall institutional objectives. The proposed framework further assumes that the environment is dynamic and may undergo changes in response to certain interventions resulting from the analysis of data. This requires the analysis techniques to be continuously evaluated and readjusted when necessary to ensure validity; the analytics processes may themselves become the subject of analysis to ensure quality of information and decisions.
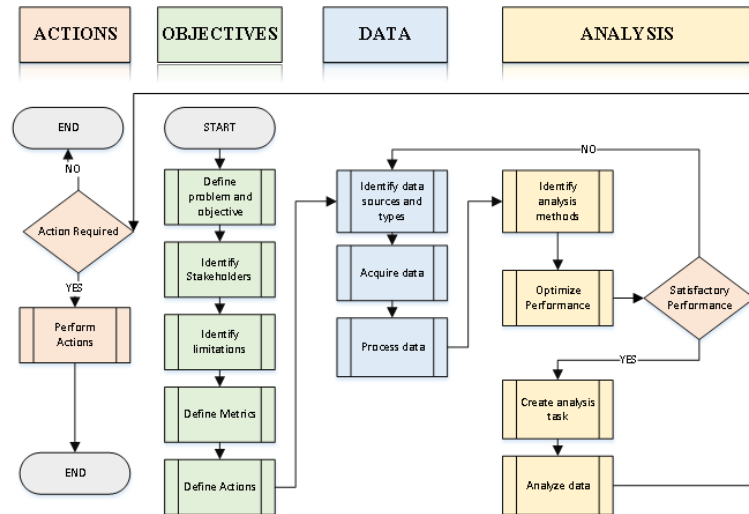
Fig. 7.2: Process flow

The process view, depicted in Figure 7.2, summarizes the procedural layer and shows the process flow from the problem definition to the actions taken. Analytics is implemented to solve a specific problem or address a defined need. It commences with the needs analysis, where the challenges and key actors are identified and metrics and actions are defined. The data acquisition phase is concerned with identifying data sources and types. It is also concerned with data acquisition. In this phase, the data analysis methods and techniques are identified and the data are analyzed. The results are the inputs to the actions phase, which is concerned with putting decisions into actions. In the next section, the framework components are discussed in greater detail followed by an application of the framework, using academic integrity as an example.

7.2.1 *Organizational needs and objectives*

The process of developing an analytics task commences with identifying the needs and defining the objectives the analytics is expected to attain. Table 7.2 depicts the procedural steps as well as the guiding questions of the objectives phase. The assessment of needs is a key element in the framework. Analytics should be viewed as a precision tool that targets specific issues rather than presents a broad report where a plan of action is developed ad-hoc. Each identified need or objective has an owner and is mapped to a set of metrics for tracking and measurement. Each analysis output is mapped to a conditional response. The intervention should be commensurate with the problem at hand. Considering that much of the problems are constructs, it is imperative to establish construct validity and validate metrics. The methods and techniques used for identifying needs, ranking their importance, and finding appropriate interventions vary among institutions, as do the needs and objectives themselves.

Table 7.2: The objectives phase

Process Step	Guiding Questions
Define problem	What is the scope of the problem?
Identify stakeholders	Who is in charge of mitigating the problem?
Identify limitations	What are the technical, budgetary, ethical, legal, administrative, and logistical limitations?
Define metrics	How is the problem being tracked and measured?
Define actions and thresholds	What needs to be done when the problem reaches a certain level?

The role of analytics may vary among institutions. It may provide an advisory function to human experts making the decisions, or it may completely automate the decision-making process, minimizing the human involvement. Analytics may also provide reac-

tive or proactive responses to a problem. The objectives phase lays the foundation for the subsequent steps and is concerned with narrowing the scope of the analysis task, mapping stakeholders to the specific problem being solved, creating metrics, and outlining the types of interventions and conditions under which the interventions will be triggered. The scope of issues that will be addressed through the use of analytics depends on the specific institutional context. For example, a survey of 861 research papers was carried out by Bozkurt et al (2015) to classify current trends in distance education research published between 2009 and 2013 in seven peer-reviewed journals. The results suggest that much of the research examined the learners' emotional states, focusing on gauging student satisfaction and learner perception." The results suggest a keen interest in profiling learners and identifying their individual differences. Some of the variables used in the distance learning research include: perceptions, communication, age, satisfaction, academic-performance, self-efficacy, participation, collaboration, interaction, social-presence, and motivation.

After the problem and key actors are identified, an assessment of constraints that may influence the design of the analytics is performed. Its aim is to identify environmental constraints that, among others, may include technical, financial, ethical, legal, administrative and logistical issues. The monitoring of the objectives may require data that is not be readily available, is expensive, or the collection of which is subject to legal restrictions and ethical reviews. Assessing the means required for the successful execution of the analysis early on will prevent shortfalls during implementation and delivery.

7.2.2 Data

The data phase of the framework aims to collect and feed the analysis component with the relevant data. The process steps and related questions are presented in Table 7.3. Data can be broadly organized into two categories: the internal data—that the organization collects (e.g., student-generated content), and the external data—that is acquired from a third party service (e.g., IP address geolocation database). In many of the cases the analytics application will use a mixture of data sources and types. For example, event logs from computer systems are commonly used for analysis, as they are readily available and capture a variety of user behaviors. The logs may be combined with student-produced content such as forum postings and papers to perform tasks such as the sentiment analysis. Learner produced-content is readily available and constitutes a valuable resource as it may be used to support a number of academic processes, such as enforcement of academic integrity and student support. Textual data can also be used for personality profiling (Argamon et al, 2005; Noecker et al, 2013), which has a variety of applications in the learning environment.

Data are often stored in a variety of formats by different systems, which requires a tailored approach. After the data are acquired and before it can be analyzed, the data need to undergo a pre-processing step which filters out the unnecessary parts. For example, running a sentiment analysis of the course reviews does not involve information about the type of the web browser used to post messages and therefore, it may be safely removed. The steps of data acquisition and processing are technical; therefore, the end users of the analytics system do not need to interact with the raw data, but only access the outcomes of the analysis.

Table 7.3: The data phase

Process Step	Guiding Questions
Identify data sources and types	What is the data source?
Acquire data	How can the data be retrieved?
Process data	What parts of the data are irrelevant or redundant?

7.2.3 Analysis

The analysis phase of the framework is concerned with identifying the methods and techniques for analyzing the data and carrying out the analyses. The process steps and related questions are presented in Table 7.4. The analyses can be conducted using statistical and machine learning techniques. While the former is often used in quantitative research, the latter is a growing field of computer science concerned with computational theories and techniques for discovering patterns in data. A key advantage of machine learning over statistical techniques is that machine learning algorithms learn from data without being specifically programmed for each task. The analyses can be organized into four categories: diagnostic, descriptive, predictive, and prescriptive (Herschel et al, 2015). The first two categories deal with past or present events, while the latter two categories attempt to glimpse into the future.

The analysis techniques are expected to demonstrate the validity of the selected approaches before the results are translated into actions. To this end, the proposed framework provides an optimization loop between the data and analysis phases that enables finding the optimal balance between the data and the computational approach. It also facilitates the fine-tuning of the data selection and processing, and testing of the computational methods prior to deployment into the production environment. The computational performance can and should be quantitatively assessed; a number of metrics are available

to estimate the performance (Sokolova and Lapalme, 2009; Fawcett, 2006). Data may be analyzed on-demand or continuously depending on the stakeholder requirements, computational resources, and the type of data used. For example, enrollment data may be analyzed every time the new data becomes available, whereas a comparison of the final grades between two courses is performed upon request. The process of running the analysis should be friendly enough for non-experts to use and communicate to others.

Table 7.4: The analysis phase

Process Step	Guiding Questions
Identify analysis methods	What is the analysis technique?
Optimize performance	Are the results valid?
Create analysis task	Can a non-expert operate it?
Analyze data	What are the results?

The output of the analysis process is the input to the decision making process where actions are congruent with the stakeholder requirements. In some cases, the analysis results may be presented in the form of a report issued to select stakeholders who then carry out the actions; in other cases, actions may be automated.

7.2.4 Actions

The actions phase of the framework is concerned with translating the analysis results into interventions. Actions are defined in the needs phase and constitute measurable and conditional responses to the output of the data analysis. They aim to answer the question—"What if?" For example, what would happen if the student grades fell beyond a pre-defined threshold level? Analytics may serve as an advisor to the stakeholders who

act upon the received information, or may serve as an actor that carries out the decisions automatically without human involvement. The process steps of the action phase and corresponding questions are presented in Table 7.5.

Analytics systems may grow with time and add multiple objectives, analyses and interventions. Some analytics tasks will only yield the reports to the stakeholders, who will then take the necessary actions (e.g., conduct academic integrity review), while other tasks will trigger automatic actions (e.g., produce a list of recommended courses.) When the actions are automated, the stakeholders may still monitor the quality of the task execution and intervene if necessary. Interventions, akin to data analyses and data structures, are all context-specific items. For example, an analytics task that performs automatic screening of plagiarism in student-produced content may be configured to produce a warning message for the student's eyes only, or it may trigger a report for the instructor to investigate the issue.

Table 7.5: The action phase

Process Step	Guiding Questions
Action required	Are the conditions satisfied for taking action?
Perform action	Was the intervention delivered?

7.3 Application of the framework

This section describes application of the framework to support implementation of the data-driven academic integrity approach presented in Chapter 6.

Analytics serves an advisory function to the course instructors. Implementation of data-driven is a team effort and the instructors will rely on support from the IT and research teams. The development teams need to work jointly with the experts to customize analysis processes and provide any necessary technical training on the use of analytics and interpretation of the results. Because the assessment is done algorithmically, this approach provides a more efficient alternative to having human proctors physically observe students. Each procedural step and corresponding outputs are elaborated in Tables 7.6-7.9.

7.4 Summary

Because the proposed framework views analytics as a sum of parts, institutions may pick and choose the analytics tasks that are congruent with their needs and objectives. For example, emotive state identification algorithms can be added to the academic integrity framework, or a content-level analysis may be excluded in some courses. The piecemeal approach promotes scalability and efficiency by investing in only what is required to solve a particular issue.

Analytics is a tool that helps to keep a grip on the business and academic environments; however, in itself, it should not be considered a sole remedy to problems that arise, because solutions entail actions. It is important to highlight a distinction between effectiveness of the data analysis techniques and that of the administrative processes. If the stakeholders are reluctant to accept and act upon the information conveyed by the data analysis, the problem that the analytics was implemented to address will not be resolved. In the case of automated actions, such as recommender systems, the stakeholders

Table 7.6: The objectives phase

Process Step	Output	Delegated to
Define problem	The need to provide academic integrity is stipulated by the accreditors. The objective is to provide assurance that students are who they say they are and that they did the work they say they have done. The context of the analytics task implementation is a graduate-level research methods course delivered online by a European university. The course comprises four written assignments.	Admin., Faculty
Identify stakeholders	The main stakeholders in this process are: the students, whose assignments will be analyzed; the instructors who are the owners of the analytics task; and academic administrators who control the quality of the educational process. The analytics task is developed by the research team in cooperation with the IT department who will provide the necessary access to the student data, facilitate the development, and provide training to users.	Admin., Faculty
Identify limitations	Data collection procedures require compliance with the university policies. Written assignments are the principal means for course evaluation. The analyses will be conducted after the second assignments, as a minimum of two assignments is required to perform classification tasks. There is no LMS integration and analyses and reporting are performed externally to the LMS. Classification results cannot be calculated to a certainty and instructor review will be required in cases of misclassification. Training on how to use the analytics task will be required and provided to the instructors by the development team.	Admin, Faculty, IT
Define metrics	This task is concerned with the measurement of the patterns of language use, such as the frequency of word pairs and triplets compared across assignment pairs of each student. Cases of misclassification will be flagged for the instructor intervention.	Admin, Faculty, IT
Define actions	The analysis will produce a classification report. Any instance of misclassification will be considered a case of potential misconduct and trigger a manual review by the instructor. Any remedial actions will be taken at the instructor's discretion.	Admin, Faculty

Table 7.7: The data phase

Process Step	Output	Delegated to
Identify data sources and types	The data source for this analytics task will be limited to the textual content produced by students. The students submit their assignments through the LMS, which stores the files in their native formats. The file-naming convention includes the student's name, number of the learning activity, a part of the original title, and a timestamp, which will be used for identification and labeling of the document authors. The course comprises four learning activities; therefore, four files per student will be retrieved.	Faculty, IT
Acquire data	The IT team will facilitate access to the LMS data store from where the documents will be retrieved.	IT
Process data	The assignments are produced in a variety of formats and include Word document, the portable document format, and Latex, which requires text to be parsed from each file type prior to the analysis. Additional pre-processing steps such as the removal of noise, which includes symbols, names, headers, footers, and direct citations, also need to be performed.	IT

Table 7.8: The analysis phase

Process Step	Output	Delegated to
Identify analysis methods	The analysis will follow the protocol outlined in Amigud et al (2016). The data will be retrieved, processed and analyzed to test the techniques and adjustments will be made to make the process more efficient and effective. Accuracy measure is used to assess classification performance.	Research, IT
Optimize	To improve performance, the protocol is amended to include feature selection as outlined in (Amigud et al, 2017a). Feature weighing is employed and limited to the top 300 features. The feature set comprises bigrams with stop words preserved. Texts are split into 500 word chunks.	Research
Create analysis task	Analyses are conducted using Scikit-Learn (Pedregosa et al, 2011) machine learning library in the Python language. An intuitive graphical interface is created (Amigud et al, 2017b) to allow the faculty to perform analysis on-demand.	IT
Analyze data	The assignments are passed on to the analytics engine. The outcome of the analysis is a report that identifies students whose work requires a closer look.	Faculty

Table 7.9: The actions phase

Process Step	Output	Delegated to
Action required	Upon examining the analysis report and the students' work, the instructor is satisfied with the outcome.	Faculty
Perform action	No additional administrative actions will be taken at this time.	Faculty

need to maintain quality control and ensure that analytics is effective. The system can yield highly accurate predictions and a vast array of reports, but if the stakeholders fail to act upon the information and resolve the problem, the analytics is not going to make the problem disappear.

The relationship between information and actions is of paramount importance. The success of an analytics strategy is predicated on a clear understanding of what the needs are, how they are tracked and measured, and what actions should be taken to address them. The type of information monitored through the use of analytics and the way it is processed depends on the institutional context. This suggests that implementation of analytics is the process of adaptation and customization. One should be careful when choosing out-of-the-box analytics products and ensure that the metrics reflect the institutional context at hand. Analytics is in a dynamic state, because the needs and contextual limitations are subject to change, and it requires continuous reevaluation. As new types of data emerge, so do new computational techniques. "Install it and forget it" may not be the best strategy to follow— when it comes to analytics tools— as it will inevitably lead to poor decisions. Reevaluation of analytics is beneficial not only for the improvement of the quality of information and decisions, but also for compliance with legal requirements and ethical standards that themselves are subject to change.

Conclusion

This thesis is predicated on the idea that provision of academic integrity should not be invasive, disruptive, inconvenient, or logistically challenging. The literature review uncovered the gaps in traditional approaches to academic integrity, and highlighted areas of opportunity for developing an alternative to the traditional academic integrity approaches that aim to provide greater accessibility, convenience, unobtrusiveness, privacy, and robustness of identity and authorship assurance.

To address these gaps, a one-tiered approach to identity and authorship assurance was designed, developed, and evaluated using real-world student data. The results suggest that its performance exceeds that of instructors, and that the majority of participants positively perceived the approach. Instructor validation has its merits and in some cases may provide the first line of defense against academic misconduct; however, the data suggests that the ability of instructors to identify potential academic misconduct is inferior to that of the computational methods. Computational approach enables automation of identity and authorship assurance tasks, calling for an instructor only in cases where manual intervention is required. Because the analyses are conducted on the background— without a need for interaction—students can learn their way, without suffering restrictions and without subjecting their private learning spaces to monitoring and control. The proposed

approach also enables a greater level of user privacy as it eliminates the need to share access to students' personal computers or the need to share physical biometrics with a third party in order to complete identity verification (Levy et al, 2011).

The advantages of e-assessments over traditional pen and paper exams have been noted, and some maintain that traditional exams are being gradually replaced by electronic assessments (Ramu and Arivoli, 2013). Many of the institutions with distance programs employ authentic assessment such as assignments, projects, and portfolios (Bailie and Jortberg, 2008) in lieu of secure exams. These type of assessments are often delivered continuously. However, continuous assessment often entails more frequent interaction between the student and instructor, and evaluation and feedback follows each unit of learning. To address these issues, the proposed method and system provide the necessary means for both the user-level analysis, to map learner identities with their academic work, and also the content-level analysis, to simplify the review of the student-produced content. The contributions of this research and future directions are discussed in the following sections.

Thesis contributions

This thesis is based on six publications, each of which contributes to the fulfillment of the thesis goals.

- Amigud, A., Arnedo-Moreno, J., Daradoumis, T., Guerrero-Roldan, A.-E. (2017). Using Learning Analytics for Preserving Academic Integrity. *The International Review of Research in Open and Distributed Learning, 18*(5). doi:10.19173/irrodl.v18i5.3103 (JCR IF=1.732, Q1).

- Amigud, A., Daradoumis, T., Arnedo-Moreno, J., Guerrero-Roldan, A.-E. (2018). A Procedural Learning and Institutional Analytics Framework. in S. Caballe & J. Conesa, (Eds.), *Software Data Engineering for Network eLearning Environments: Learning Analytics and Context Awareness Services (Lecture Notes on Data Engineering and Communications Technologies)*. Berlin, Germany: Springer-Verlag GmbH.

- Amigud, A., Arnedo-Moreno, J., Daradoumis, T., Guerrero-Roldan, A.-E. (2017). A Method for Thematic and Structural Visualization of Academic Content. *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*. doi:10.1109/icalt.2017.24

- Amigud, A., Arnedo-Moreno, J., Daradoumis, T., Guerrero-Roldan, A.-E. (2017). A Robust and Non-invasive Strategy for Preserving Academic Integrity in an Open and Distance Learning Environment. *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*. doi:10.1109/icalt.2017.23

- Amigud, A., Arnedo-Moreno, J., Daradoumis, T., Guerrero-Roldan, A.-E. (2017). Open Proctor: An Academic Integrity Tool for the Open Learning Environment. *Lecture Notes on Data Engineering and Communications Technologies, 262–273*. doi:10.1007/978-3-319-65636-6_23.

- Amigud, A., Arnedo-Moreno, J., Daradoumis, T., Guerrero-Roldan, A.-E. (2016). A Behavioral Biometrics Based and Machine Learning Aided Framework for Academic Integrity in E-Assessment. *2016 International Conference on Intelligent Networking and Collaborative Systems (INCoS)*. doi:10.1109/incos.2016.16

The paper titled "Using Learning Analytics for Preserving Academic Integrity" discusses the traditional approaches to academic integrity, their advantages, and shortcomings. It introduces the notion of individual-specific patterns of language use that enable delineation of content by author. It reports on the results of experiments aimed at establishing a baseline of human performance and comparing it to that of computational methods. To the author's knowledge, this was the first study to compare instructor performance to that of the computational methods. The paper titled "A Behavioral Biometrics Based and Machine Learning Aided Framework for Academic Integrity in E-Assessment" ex-

amines the application of authorship analysis on student-produced textual data with an aim to establish a proof of concept. Two papers titled "A Robust and Non-invasive Strategy for Preserving Academic Integrity in an Open and Distance Learning Environment" and "Open Proctor: An Academic Integrity Tool for the Open Learning Environment" report on the development of a prototype application. Functionality of the prototype was expanded to include content-level analysis which was explicated in the paper titled "A Method for Thematic and Structural Visualization of Academic Content". A roadmap for framework implementation is presented in a book chapter titled "A Procedural Learning and Institutional Analytics Framework". Together, these publications share a common theme of simplifying learner assessment and enhancing academic integrity using a computational approach.

Accomplishment of thesis goals

The aim of this thesis was to develop an alternative academic integrity framework that addresses the shortcomings of the traditional, observation-based academic integrity strategies that merely collect evidence of misconduct. To this end, this thesis posed nine research questions each corresponding to an objective that addresses theoretical, empirical, or pragmatic goals.

The first objective was to review current literature on academic integrity strategies and delineate those that constitute promising approaches to providing single-tiered identity and authorship assurance. Chapter 1 did just that and provided an answer to the first research question — what academic integrity strategies can concurrently provide identity and authorship assurance in a convenient and non-invasive fashion? It identified behav-

ioral biometrics as an approach that requires no special hardware while providing concurrent identity and authorship verification. More particularly, stylometry appeared to have potential to provide non-disruptive assurance of both identity and authorship without a need for live and continuous data collection. Stylistic analysis is performed on the background, and takes advantage of the available student-produced content. The next objective was to examine the available stylometric techniques, methods and tools to gain a better understanding of how it can be integrated into an academic integrity framework and future system. Chapter 2 addressed the questions of what are the techniques for authorship analysis and how well do they perform? It provided an overview of the tools, datasets, and computational techniques for conducting authorship analysis tasks.

Moving from theory to empirical evaluation of the computational-based strategy, the objectives were to develop and evaluate a computational academic integrity framework using real world data. Chapter 3 presented a proof of concept with an aim to answer the question of whether the student-produced content could be used for providing identity and authorship assurance. The preliminary results suggested that it was possible to map learner identities to their work through analysis of textual data. However, the performance varied across learning activities that highlighted a need to refine the analysis techniques and improve performance. This became the next objective. The first part of Chapter 4 answered the question of what is the baseline of human performance? A study to establish performance of the education practitioners classifying short academic texts by author was conducted. It examined a notion by Barnes and Paris (2013) that instructors should be able to delineate students' writing styles. The second part of Chapter 4 addressed the question of how can classification accuracy be improved? A combination of ensemble techniques, lexical and syntactic feature sets, and feature weighing were employed. The

results suggest that the instructors performed at a lower rate than the computational-based methods.

Once the computational process was developed, it was time to develop a framework and turn it into a computer application that provides identity and authorship assurance. More formally, the objective was to integrate research findings into an e-assessment system and develop prototype application. To this end Chapter 5 addressed the question of what would a system and method for maintaining academic integrity in the learning environment, based on behavioral pattern analysis of learner-produced content, be like? The prototype system was modular; it enabled students to upload the content and instructors to perform administrative functions such as user and file management and run the analysis that flags cases of potential misconduct. Once developed it was interesting to examine the perceptions of instructors toward the proposed academic integrity approach. The second part of Chapter 5 examined the question of what are the instructors perceptions towards the new academic integrity approach? To this end, the application entitled OpenProctor was first showcased at the IABL conference in Toronto. Live demonstrations were conducted. The survey responses from nine survey participants suggest an interest and positive perception of the technology. Considering the modular structure of the proposed system, the next logical step was to enhance the functionality of the prototype to serve not only as an academic integrity system but as an e-assessment platform. Chapter 6 addressed the question of what can be done to make the review of written content more efficient? To this end, a method for thematic and structural visualization of textual data was developed and integrated into the prototype. The last objective was to create a framework that simplifies implementation of the data-driven approach into institutional practices. That corresponding research question was, what procedural steps are

required to integrate data-driven approaches into academic and institutional practices? These were discussed in Chapter 7.

Future research

This work expanded theoretical understanding of academic integrity issues, and also yielded pragmatic outcomes. This line of research is expected to remain important for as long as trust and honesty remain the core values of academic culture. For those who wish to continue research in the area of academic integrity, I would like to make five recommendations for future research.

- Performance baseline is a precursor to assessing effectiveness. In order to make quantitative judgments and relative comparisons between the methods and tools, the acceptable performance levels for each approach needs to be explicitly defined. The future research should focus on establishing criteria for measures and levels of service sufficient to keep the learning environment secure.

- An informed decision to choose one academic integrity strategy over another requires knowledge of how well they perform. Closing the gap in the literature entails conducting research on the effectiveness of integrity strategies both perceived (case-based) and relative (baseline).

- Much of the literature discusses theoretical, proof of concept, or pilot projects. The data on the long-term use or adaptations of academic integrity strategies is not available. The future research should examine institutional experience managing academic integrity, their success stories and challenges.

- Academic integrity strategies vary in the levels of convenience, privacy, automation, disruption, and assurance they deliver. The impact of academic integrity measures on student performance should be further examined.

Furthermore, the focus of this thesis was on the student behavior, however, academic integrity applies equally to researchers, faculty, editors, and reviewers. The scope of the future research should include the relationships of these stakeholder groups and contextual factors that influence unethical behavior.

Appendix A

Surveys and Related Documents

For each study involving participants or learner-produced data, a request was submitted to the Ethics Committee at UOC.

A.1 Evaluation of computational performance

This thesis posed a question of whether or not analysis of the student-produced content can be used for providing identity and authorship assurance. Chapters 3 and 4 address this question and propose a computational framework. A request was filed with the UOC to obtain a copy of the student-generated content. Data comprising of forum postings and continuous assessment exercises were collected from the LMS. All experiments were computational, no human interaction was required, and there was no need for participant recruitment. All student data were anonymized. The ethics committee approval letter is depicted in Figure A.1.

UOC
Universitat Oberta
de Catalunya

Evaluation of the "Computational evaluation of the writing style" project by the Ethics Committee of the UOC

Dr. Marta Aymerich Martínez, president of the Ethics Committee of the Universitat Oberta de Catalunya

CERTIFIES

That the Committee has evaluated the proposal submitted by UOC Ph.D student NIT Doctoral Programme Alexander Amigud, regarding the "**Computational evaluation of the writing style**" project, led by Atanasi Daradoumis of *Universitat Oberta de Catalunya*, and in which, he will participate by collecting and having access to research data.

AND

that regarding the "Instructor perceptions of the automated tool for continuous academic integrity screening" project

- The ability of the researchers and their collaborators, and the facilities and resources available are adequate to carry out the study.
- The established experimental protocol ensures the integrity and dignity of the participants.
- The protocol is adequate to the objectives of the study and the possible risks and discomfort for participants are adequate given the expected benefits.
- The procedure for obtaining informed consent of participants, including the information sheet, and the procedure for the recruitment of subjects are adequate.

Having met on 18th October 2016, and having considered the ethical implications concerning human experimentation and the processing of personal data, this committee APPROVES the "Computational evaluation of the writing style" project.

For the record, I sign this document in Barcelona, date of decision 18th October 2016.

Signed:

Dr. Marta Aymerich Martínez

Av. Tibidabo, 39-43
08035 Barcelona – Spain
Tel. +34 93 253 23 00
Fax +34 93 417 64 95

Fig. A.1: Student data study approval

A.2 Evaluation of human performance

In order to make quantitative judgements about the proposed academic integrity approach, a comparative baseline was required. However, the literature is sparse and does not provide a definitive answer as to how different approaches compare on performance criterion. To bridge the literature gap, a study was conducted to assess the efficacy of instructor validation—the most basic academic integrity strategy where instructors identify irregularities in the student work. Considering that instructors or teaching assistants continuously interact with students and/or their academic content through discussions and grading of the assignments, instructor validation seems to be a possible defense against cheating. Barnes and Paris, (2013) argued that "key to recognizing cheating or plagiarism is to become familiar with a student's writing style." This notion was put to the test and instructors were invited to participate in the study to classify student writings by author. The methodology and results are explicated in Chapter 4. An excerpt of the data collection instrument is depicted Figure A.2. The ethics committee approval letter is depicted in Figure A.3.

A.2.1 Invitation to participate in a study

Dear {PARTICIPANT},

My name is Alexander Amigud and I am a doctoral student in the Network and Information Technologies program at the Universitat Oberta de Catalunya. I am conducting a study that examines the perception of authorial style and aims to establish a baseline of human performance classifying short texts. This research has implications for the process

Which of the following 3 text samples were written/authored by the same person/author?

Please explain your selection.

Sample 1:

The six relevant aspects of research can be summarised as the 6Ps, including purpose, products, process, participants, paradigm and presentation. Purpose refers to the reason behind the conducted research, more specifically to the research question that the research aims at answering. The product of the research can be both, physical and non-physical. Non-physical products include the answer to the research question and additional findings, whereas physical products are the medium through which the findings are conveyed: for example a journal article, dissertation or, in case of design and creation research, a newly developed application. Process refers to the actual execution of the research and consists itself of five steps, namely identifying the research topic, developing a conceptual framework, selecting a research strategy and data-generating process, analysis of the data and lastly concluding. The fourth P refers to the research participants, which refer to those directly, as well as indirectly, involved in the research and the researcher himself. Fifthly, philosophical paradigms refer to a common view on something. These include positivism (prevalent in the U.S.), interpretivism (prevalent in Europe) and critical research. Lastly, the presentation is concerned with conveying the results to others e.g. though a thesis or conference presentation. These six aspects must be considered in any research project.

Whereas systems development is concerned with creating new software, research aims at adding to the body of knowledge. The SDLC consists of four main stages, namely analysis, design, implementation and testing. Each of these can be related to a stage in the research process. It begins with an analysis of the status quo, which, in case of systems development, refers to the analysis of the current, as well as the proposed system (if one already exists) and the creation of a requirements specification, in which the objectives for the system are defined. Similarly, the analysis stage in research would include the literature review, which defines the state of the art and a gap in the literature, which is then addressed in one or several research questions. The second stage design refers to the specification of the features of the planned software from general to more specific ones in systems development and the planning of the research approach from the general research strategy (e.g. survey or case study) to the specification of the data generating method (e.g. interview or observation) within the chosen research strategy. The third step consists of the implementation. This refers to the physical production of the system by following the design chosen in the second step. For researchers, this relates to the implementation of the chosen strategy and the gathering of empirical data through the proposed method; combined this can be considered the data analysis. Lastly, the testing stage itself consists of two components: On the one hand, the proposed system must meet the requirements defined at stage one and the conducted research should answer the research questions. But it is just as important that the proposed product appeals to the participants.

Sample 2:

The six relevant aspects of research can be summarised as the 6Ps, including purpose, products, process, participants, paradigm and presentation. Purpose refers to the reason behind the conducted research, more specifically to the research question that the research aims at answering. The product of the research can be both, physical and non-physical. Non-physical products include the answer to the research question and additional findings, whereas physical products are the medium through which the findings are conveyed: for example a journal article, dissertation or, in case of design and creation research, a newly developed application. Process refers to the actual execution of the research and consists itself of five steps, namely identifying the research topic, developing a conceptual framework, selecting a research strategy and data-generating process, analysis of the data and lastly concluding. The fourth P refers to the research participants, which refer to those directly, as well as indirectly, involved in the research and the researcher himself. Fifthly, philosophical paradigms refer to a common view on something. These include positivism (prevalent in the U.S.), interpretivism (prevalent in Europe) and critical research. Lastly, the presentation is concerned with conveying the results to others e.g. though a thesis or conference presentation. These six aspects must be considered in any research project.

Whereas systems development is concerned with creating new software, research aims at adding to the body of knowledge. The SDLC consists of four main stages, namely analysis, design, implementation and testing. Each of these can be related to a stage in the research process. It begins with an analysis of the status quo, which, in case of systems development, refers to the analysis of the current, as well as the proposed system (if one already exists) and the creation of a requirements specification, in which the objectives for the system are defined. Similarly, the analysis stage in research would include the literature review, which defines the state of the art and a gap in the literature, which is then addressed in one or several research questions. The second stage design refers to the specification of the features of the planned software from general to more specific ones in systems development and the planning of the research approach from the general research strategy (e.g. survey or case study) to the specification of the data generating method (e.g. interview or observation) within the chosen research strategy. The third step consists of the implementation. This refers to the physical production of the system by following the design chosen in the second step. For researchers, this relates to the implementation of the chosen strategy and the gathering of empirical data through the proposed method; combined this can be considered the data analysis. Lastly, the testing stage itself consists of two components: On the one hand, the proposed system must meet the requirements defined at stage one and the conducted research should answer the research questions. But it is just as important that the proposed product appeals to the participants.

Sample 3:

The six relevant aspects of research can be summarised as the 6Ps, including purpose, products, process, participants, paradigm and presentation. Purpose refers to the reason behind the conducted research, more specifically to the research question that the research aims at answering. The product of the research can be both, physical and non-physical. Non-physical products include the answer to the research question and additional findings, whereas physical products are the medium through which the findings are conveyed: for example a journal article, dissertation or, in case of design and creation research, a newly developed application. Process refers to the actual execution of the research and consists itself of five steps, namely identifying the research topic, developing a conceptual framework, selecting a research strategy and data-generating process, analysis of the data and lastly concluding. The fourth P refers to the research participants, which refer to those directly, as well as indirectly, involved in the research and the researcher himself. Fifthly, philosophical paradigms refer to a common view on something. These include positivism (prevalent in the U.S.), interpretivism (prevalent in Europe) and critical research. Lastly, the presentation is concerned with conveying the results to others e.g. though a thesis or conference presentation. These six aspects must be considered in any research project.

Whereas systems development is concerned with creating new software, research aims at adding to the body of knowledge. The SDLC consists of four main stages, namely analysis, design, implementation and testing. Each of these can be related to a stage in the research process. It begins with an analysis of the status quo, which, in case of systems development, refers to the analysis of the current, as well as the proposed system (if one already exists) and the creation of a requirements specification, in which the objectives for the system are defined. Similarly, the analysis stage in research would include the literature review, which defines the state of the art and a gap in the literature, which is then addressed in one or several research questions. The second stage design refers to the specification of the features of the planned software from general to more specific ones in systems development and the planning of the research approach from the general research strategy (e.g. survey or case study) to the specification of the data generating method (e.g. interview or observation) within the chosen research strategy. The third step consists of the implementation. This refers to the physical production of the system by following the design chosen in the second step. For researchers, this relates to the implementation of the chosen strategy and the gathering of empirical data through the proposed method; combined this can be considered the data analysis. Lastly, the testing stage itself consists of two components: On the one hand, the proposed system must meet the requirements defined at stage one and the conducted research should answer the research questions. But it is just as important that the proposed product appeals to the participants.

○ Text samples 1 and 2

○ Text samples 2 and 3

○ Text samples 1 and 3

● All 3 text samples were written by the same author

○ Each text sample was written by a unique author

○ No answer

Please enter your comment here:

Same text.

Fig. A.2: Stylistic perception survey question

Evaluation of the "Empirical evaluation of instructor validation effectiveness" project by the Ethics Committee of the UOC

Dr. Marta Aymerich Martínez, president of the Ethics Committee of the Universitat Oberta de Catalunya

CERTIFIES

That the Committee has evaluated the proposal submitted by UOC Ph.D student NIT Doctoral Programm Alexander Amigud, regarding the "**Empirical evaluation of instructor validation effectiveness**" project, led by Atanasi Daradoumis of Universitat Oberta de Catalunya, and in which, he will participate by collecting and having access to research data.

AND

that regarding the "Empirical evaluation of instructor validation effectiveness" project

- The ability of the researchers and their collaborators, and the facilities and resources available are adequate to carry out the study.
- The established experimental protocol ensures the integrity and dignity of the participants.
- The protocol is adequate to the objectives of the study and the possible risks and discomfort for participants are adequate given the expected benefits.
- The procedure for obtaining informed consent of participants, including the information sheet, and the procedure for the recruitment of subjects are adequate.

Having met on 20th September 2016, and having considered the ethical implications concerning human experimentation and the processing of personal data, this committee APPROVES the "Empirical evaluation of instructor validation effectiveness" project.

For the record, I sign this document in Barcelona, date of decision 20th September 2016.

Signed:

Dr. Marta Aymerich Martínez

Av. Tibidabo, 39-43
08035 Barcelona – Spain
Tel. +34 93 253 23 00
Fax +34 93 417 64 95

Fig. A.3: Human performance study approval

of learning and teaching, academic integrity, and technology selection. I have created a short online survey consisting of multiple-choice questions asking to identify short text passages (excerpts from student assignments) that look stylistically similar and are perceived to belong to the same author. I was wondering if you would be interested in taking part in this study. This should not take more than 30 minutes to complete. I very much appreciate your help. The survey is located here: {SURVEYURL}

Sincerely,

Alexander Amigud

About the study: Writing style is thought to be author-specific, but is it possible to identify authors of short texts without the use of computer tools? This study aims to examine the perception of authorial style and establish a baseline of human performance classifying short texts. To this end, an online survey comprising of excerpts from 5 student papers, randomly distributed over 5 questions, was created. You are invited to participate in this study and try to predict which text samples were written by the same person. This study has implications for the process of learning and teaching, academic integrity, and technology selection. Students and instructors continuously interact with textual content. Instructors have a great level of exposure to a variety of textual content produced by the same group of authors (students), throughout the learning cycle. The online survey will also include introductory questions such as teaching experience, number of courses taught, and the number of assignments graded.

There will be no personal information collected during the survey. All data will be securely stored, encrypted, password protected and deleted upon project completion.

Please note that participation in this study is voluntary. You may withdraw from the study at any time. If for any reason you decide to withdraw, your data will not be used and will be deleted.

The final research paper will be made publicly available. A copy of the paper will be provided to you. The research may also be used for future publication of presentation in academic or professional journals and conferences. This study has been reviewed by the Universitat Oberta de Catalunya Research Ethics Board. Should you have any comments or concerns regarding your treatment as a participant in this study, please contact the Office of Research Ethics at (+34) 93-253-2300 or by email at comite_etica@uoc.edu.

Alexander Amigud

Researcher

Email: <EMAIL>

If you have any questions, please do not hesitate to contact me or my research supervisors:

Dr. Atanasi Daradoumis

Universitat Oberta de Catalunya

IT, Multimedia and Telecommunications Department

Email: adaradoumis@uoc.edu

Dr. Joan Arnedo Moreno

Universitat Oberta de Catalunya

IT, Multimedia and Telecommunications Department

Email: jarnedo@auoc.edu

_____-

If you do not wish to participate in this survey and don't want to receive any more invitations please click the following link: <LINK>

A.3 Technology perception survey

After the prototype application was developed, it was important to examine how the instructors perceive the idea of using analytics-based approach for providing academic integrity. Chapter 5 discussed this in some detail. An excerpt of the survey instrument is depicted in Figures A.4 and A.5. The approval of the ethics committee is presented in Figure A.6.

A.3.1 Invitation to participate in a survey

Dear {PARTICIPANT},

My name is Alexander Amigud and I am a doctoral student in the Network and Information Technologies program at the Universitat Oberta de Catalunya. I am conducting a study that examines instructors' perceptions and attitudes towards a novel analytics-based academic integrity tool. I was wondering if you would be interested in taking part in this study. This should not take more than 30 minutes to complete. I very much appreciate your help. The survey is located here: {SURVEYURL}

Sincerely,

Alexander Amigud

Fig. A.4: Technology perception survey

About the study: The credibility of higher education is challenged by the rapid growth of ubiquitous computing that bridges the time and space gap between learners and instructors and creates new opportunities for academic misconduct. This study introduces a new type of academic integrity tools and aims to examine the perceptions towards this type of technology. The study comprises: a demonstration video, a whitepaper of the

Fig. A.5: Technology perception survey (cont'd)

novel approach to using analytics for providing academic integrity, and a 10 question survey. It should not take more than 30 minutes to complete the survey.

You are invited to participate in this study which has implications for the process of learning and teaching, academic integrity, and technology selection. There will be no personal information collected during the survey. All data will be securely stored, encrypted, password protected and deleted upon project completion.

⊐UOC
Universitat Oberta
de Catalunya

Evaluation of the "Instructor perceptions of the automated tool for continuous academic integrity screening" project by the Ethics Committee of the UOC

Dr. Marta Aymerich Martínez, president of the Ethics Committee of the Universitat Oberta de Catalunya

CERTIFIES

That the Committee has evaluated the proposal submitted by UOC Ph.D student NIT Doctoral Programm Alexander Amigud, regarding the **"Instructor perceptions of the automated tool for continuous academic integrity screening"** project, led by Atanasi Daradoumis of Universitat Oberta de Catalunya, and in which, he will participate by collecting and having access to research data.

AND

that regarding the "Instructor perceptions of the automated tool for continuous academic integrity screening" project

–   The ability of the researchers and their collaborators, and the facilities and resources available are adequate to carry out the study.
–   The established experimental protocol ensures the integrity and dignity of the participants.
–   The protocol is adequate to the objectives of the study and the possible risks and discomfort for participants are adequate given the expected benefits.
–   The procedure for obtaining informed consent of participants, including the information sheet, and the procedure for the recruitment of subjects are adequate.

Having met on 20th September 2016, and having considered the ethical implications concerning human experimentation and the processing of personal data, this committee APPROVES the "Instructor perceptions of the automated tool for continuous academic integrity screening" project.

For the record, I sign this document in Barcelona, date of decision 20th September 2016.

Signed:

Dr. Marta Aymerich Martínez

Av. Tibidabo, 39-43
08035 Barcelona – Spain
Tel. +34 93 253 23 00
Fax +34 93 417 64 95

Fig. A.6: Technology perception study approval

Please note that participation in this study is voluntary. You may withdraw from the study at any time. If for any reason you decide to withdraw, your data will not be used and will be deleted.

The final research paper will be made publicly available. A copy of the paper will be provided to you. The research may also be used for future publication of presentation in academic or professional journals and conferences. This study has been reviewed by the Universitat Oberta de Catalunya Research Ethics Board. Should you have any comments or concerns regarding your treatment as a participant in this study, please contact the Office of Research Ethics at (+34) 93-253-2300 or by email at comite_etica@uoc.edu.

Alexander Amigud

Researcher

< Email>

If you have any questions, please do not hesitate to contact me or my research supervisors:

Dr. Atanasi Daradoumis

Universitat Oberta de Catalunya

IT, Multimedia and Telecommunications Department

Email: adaradoumis@uoc.edu

Dr. Joan Arnedo Moreno

Universitat Oberta de Catalunya

IT, Multimedia and Telecommunications Department

Email: jarnedo@auoc.edu

———————————————————-

If you do not wish to participate in this survey and don't want to receive any more invitations please click the following link: <LINK>

Appendix B

Identity and Authorship Assurance Strategies

B.1 Method

Chapter 1 reviewed the research on a number of identity and authorship assurance approaches employed to enhance integrity of assessment activities. An integrative review methodology was employed (Cooper, 1984). The aim of this review was to generalize past research findings on identity and authorship assurance strategies in the academic setting. While no search can be 100% complete, a 16 year period offers a reasonable profile of the state of the art of academic identity and authorship assurance strategies. The search was performed using different terms to describe the same concept. The search was applied to the following databases: ISI Web of Knowledge, EBSCO Electronic Journals Service (EJS), Elsevier Science Direct , Scopus, Springer link, Sage journals, Taylor & Francis, Education Resources Information Center (ERIC), and Educause. The search for literature on identity and authorship assurance strategies was limited to the period between 2000 and 2016. The keywords included: (a) e-learning security; (b) learner authentication; (c) student verification; (d) learner identification; (e) testing and assessment; (f) plagiarism detection; (g) proctoring; (h) academic cheating; (i) academic misconduct; and (j) academic dishonesty.

Additional material was selected from reference lists of articles collected from the primary search. The study requirements for the literature review included: (a) peer-reviewed journal articles; (b) books; (c) conference proceedings; and (d) white papers that discuss the issue of identity or authorship assurance. The titles and abstracts of the identified literature were screened for relevance to the research questions and retrieved if deemed relevant. A total of 173 articles were initially reviewed; however, only 53 studies met the inclusion criteria. Studies were selected if they met the following inclusion criteria:

1. Published between years 2000 and 2016.

2. Published in English.

3. Discuss a method or the need of a method for identity or authorship assurance in the academic setting.

4. Have implications for security or academic integrity.

The results are presented in Table B.1. The studies are listed in alphabetical order and include: (a) the author's name and year of publication; (b) purpose; (c) main themes; (d) methodology; and (e) findings.

Table B.1: Summary of literature review

Author/Year	Purpose	Main themes	Methodology	Findings
Aceves and Aceves (2009)	To discuss the implication of the Higher Education Opportunity Act (HEOA) of 2008 as it relates to continuing and higher education.	Learner authentication, identity assurance, academic misconduct, biometrics, instructor validation, password authentication, web cameras, computer feature restrictions, proctoring, policy, legal requirements.	Position article.	There are technology-based and policy based strategies for addressing learner identity issue. The learner identification and authentication is an issue that is of concern across the higher education sector.

Amigud (2013)	To examine measures universities with large distance education programs employ to align identities of learners with the academic work they do, as well as the perceived effectiveness of and challenges and barriers to their implementation.	Identity and authorship assurance.	Survey. Multiple case study, 5 academic administrators of universities with large distance programs.	The results suggest that learner authentication strategies vary in the level of assurance. The higher degree of assurance entails higher administrative overhead, and lower accessibility.
Apampa et al (2010)	To review the existing user security model of summative assessment and to propose a model of assessment that incorporates multimodal biometrics.	Identity assurance, multimodal biometrics, identity fraud, secure assessment, proctoring.	Position article.	Authors stress the importance of learner authentication and suggest that identity assurance could be maintained through a combination of biometric applications and remote proctoring.
Auernheimer and Tsai (2005)	To describe a framework for identifying learners through a combination of password and biometrics authentication.	Identity assurance, unimodal biometrics, fingerprint, authorization, authorship assurance through proctoring, enrollment.	Theoretical framework. Prototyping a biometrics-based identification system at California State University, Fresno Digital Campus.	The study suggests that a combination of password and biometric authentication may delivery user identification with a minimal disruption to the existing internal processes.
Bailie and Jortberg (2009)	To pilot test learner identification through challenge questions and examine students' perceptions of this method.	Identity assurance, challenge questions, credibility.	Survey of online students during 2008–2009 at The National American University. Conducted 169 successful identity verifications.	Survey results suggest that students may prefer the challenge question verification to a proctor.
Beasley (2014)	To examine perceptions of students reported for an act of academic dishonesty regarding measures of prevention of such misconduct.	Academic misconduct, plagiarism, methods of prevention.	Survey of 298 undergraduate students enrolled in an academic misconduct remediation class at Michigan State University from spring 2010 to summer 2011.	The two most common reasons for cheating reported by students were ignorance of the consequences of cheating and ignorance of rules.
Barnes and Paris (2013)	To review academic integrity strategies employed by the distance instructors at Lamar University.	Identity assurance, academic integrity, policy, instructor validation, proctoring, timed exams.	Survey of 120 online instructors at Lamar University of Southeast Texas. Approximately 30 instructors completed the survey.	Instructional design measures may minimize academic misconduct. Almost 2/3 of participants consider challenge question authentication ineffective.
Bejarano et al (2013)	To present a framework for detecting plagiarism in programming assignments.	Plagiarism detection, source code.	Theoretical framework, experiment, prototype.	Performance results are promising. The proposed approach simplifies and automates screening of the assignments.
Bruhn et al (2003)	To stress the importance of access control and identity management systems.	Data security, privacy, identity assurance, authorization, standardization.	Position article.	Standardization of authentication may enable students to access various resources across institutions.

Case and Cabalka (2009)	To pilot test method of secure remote examination using Kryterion software.	Identity assurance, proctoring, video monitoring, keystroke dynamics, biometrics.	Pilot study of Kryterion remote proctoring at Western Governors University. Survey and Quantitative data were collected. 112 test administrations: 57 were administered onsite and via webcam and 75 onsite only.	The results suggest that remote proctoring may serve as an alternative to traditional proctoring.
Chiesl (2007)	To discuss practical strategies aimed at promoting academic integrity and assess their effectiveness.	Stress reduction, learner centered security, instructional design, policy, plagiarism detection.	Anonymous online survey conducted over a 3-year period. 149 students reported their online learning experience.	The results suggest that 70% of surveyed students perceive that the pragmatic method promotes academic integrity whereas 3% of students think that students will be more prone to cheat. 81% of the students reported taking each exam two to four times.
Chiranji et al (2011)	To propose a secure online examination method using group cryptography, remote monitoring and packet filtering.	Data encryption, proctoring, authorization, collusion prevention, honor code, Authorship assurance, IP monitoring.	Theoretical framework.	The framework is promising. Further research is needed to assess the validity of the proposed model.
Clarke et al (2013)	To propose a model of remote electronic invigilation of formal assessment.	Identity assurance, authorship assurance, proctoring, multimodal biometrics, continuous authentication.	Theoretical framework, prototype.	Continuous authentication for the purposes of exam invigilation is possible.
Cluskey Jr et al (2011)	To discuss an alternative to proctored examinations.	Proctoring, authorship assurance, timed exams, instructional design, computer feature restrictions.	Theoretical framework.	Authors argue that the exam proctoring cost may exceed any potential benefits. Academic integrity may be promoted through instructional design. Identity assurance may be provided through password or biometric authentication.
Culwin (2008)	To report a four-year longitudinal study on the extent of non-original content found in projects. The aim of the study was to reduce the amount of non-original content in works submitted by students.	Plagiarism detection, policy, automation, authorship assurance.	Longitudinal quantitative study of 899 screened computing projects submitted by the final-year undergraduate students enrolled in computing degree. form 2003–to 2006 at London South Bank University (LSBU), London, U.K.	The findings suggest that the adoption of an academic integrity policy that stresses education, prevention and penalty coupled with consistent application of content screening may contribute to a reduction in plagiarism rates.
Fendler and Godbey (2015)	To propose an exam design that increases costs and minimizes benefits of copying answers on multiple-choice tests.	Academic misconduct, multiple-choice test cheating, methods of prevention, instructional design, authorship assurance.	Theoretical framework.	The framework is promising. Further research is needed to assess the validity of the proposed model.

Fiedler and Kaner (2010)	To examine effectiveness of plagiarism detection tools TurnitIn and My-DropBox as well as their perceived effectiveness by the Deans.	Plagiarism detection, instructor validation.	Experiment and survey of 954 Deans of: Arts and Sciences (37%); Education (22%); Graduate Studies (22%); Engineering (11%); Business (2%); Students (1%); Law (0.2%); and Other (15%).	The study suggests that plagiarism detection tools have blind spots.
Foster et al (2009)	To describe the results of Kryterion software pilot implementation at Pennsylvania State University World Campus.	Identity assurance, keystroke dynamics, biometrics, remote proctoring, test security, feature restrictions.	Case study. In 2008 Kryterion's secure testing was pilot tested in two courses at Pennsylvania State University World Campus. A total of 29 students completed 89 exams.	The pilot results suggest that remote monitoring and authentication through keystroke dynamics may offer an alternative to traditional proctoring.
Frank (2010)	To conduct a comprehensive risk analysis of dependable distributed testing.	Test integrity, identity assurance, biometrics, continuous authentication, feature restrictions, proctoring, data security, password sharing.	Theoretical framework.	Test integrity needs to address7 types of risks: data security, identity verification, computer misuse, access to external resources, collusion or collaboration, test sharing, and electronic warfare.
Gao (2010)	To describe the commonly used academic integrity strategies and to provide a brief overview of 3 commercial proctoring products.	Identity assurance, IP monitoring, authentic assessments, credibility, password sharing, biometrics, continuous authentication, plagiarism detection.	Review.	Password-based authentication is inadequate in providing identity and authorship assurance. Continuous identity verification using biometrics during the exam may cause inconvenience for learners.
Gao (2012)	To review common academic integrity strategies and propose a method of collusion detection using IP monitoring.	Collusion detection, IP monitoring.	Experiment. Analyzed IPs of 13 students in 4 exams.	The results suggest that IP monitoring may detect potential instances of collusion.
Gregory and Strukov (2002)	To assess limitations of TurnitIn, the plagiarism detection tool.	Plagiarism detection.	Experiment consisted of 8 conditions.	The results suggest that TurnitIn has limitations in its ability to detect duplicate content.
Gullifer and Tyson (2014)	To examine students' understanding of plagiarism and their familiarity with academic integrity policy at CSU.	Plagiarism, policy.	Survey of 3405 students from Charles Sturt University (CSU) in Australia.	Findings suggest that only (52%) of the 3405 participants had read the academic integrity policy.
Heckler et al (2013)	To examine whether or not TurnItIn the plagiarism detection tool can serve as a deterrent to cheating behavior.	Plagiarism detection, authorship assurance.	Experiment, two groups 360 and 304 participants. The first group was unaware of instructor using TurnItIn, whereas the second group was.	Group that was made aware of plagiarism screening, exhibited lower incidence of plagiarism.

Hoshiar et al (2014)	To examine faculty perceptions of the effectiveness of learner authentication in online learning at community colleges.	Identity assurance, instructor validation.	Quantitative survey of 100 online faculty members from the California community colleges system.	The findings suggest the online faculty are aware of the importance of learner authentication as well as possible attempts of identity fraud. However, there was a significant gap between what faculty perceived as important for ensuring authentication and institutional practices.
Irfan et al (2009)	To examine usefulness and robustness of an alternative face recognition invigilation approach based on the discrete cosine transformation and kernel discriminant analysis.	Identity assurance, unimodal biometrics, face recognition, proctoring, security.	Experiment using 400 images of 40 subjects.	The experiment confirms usefulness and robustness of the alternative approach to face recognition.
Jortberg (2010)	To report experiences of students, faculty and administrators on using Acxiom identify-X challenge question system, as well as discuss a framework for assessing identity assurance risks.	Identity assurance, challenge questions.	Position paper. Survey of 85 students at Sullivan University.	Acxiom users reported no privacy or implementation issues. 91% of participants preferred identity verification through challenge questions to other methods of identification.
Jiao (2011)	To discuss privacy issues and propose a framework for privacy information protection within e-learning environment.	Privacy, data security.	Theoretical framework.	A combination of policy and technology may provide a greater security within the learning environment.
Kang and Kim (2015)	To discuss security and privacy considerations for e-learning courses and propose a new method of continuous authentication using biometrics.	Identity assurance, biometrics, face recognition, iris scan, continuous authentication.	Theoretical framework.	The framework is promising. Further research is needed to assess the validity of the proposed method.
Kermek and Novak (2016)	To propose a method for screening plagiarism of source code in academic assignments.	Plagiarism detection, source code.	Theoretical framework, experiment, prototype.	The framework shows promising results. Manual review is required to control for false positives.
Levey and Maynard (2011)	To evaluate student acceptance of authentication through Biometric Signature ID.	Identity assurance, biometric signature, privacy, continuous authentication, password sharing.	Pilot Study. Houston Community College and BiosigID pilot project. 140 students completed the identity enrollment process. 58 students completed the survey.	All participating students were able to enroll and complete authentication. Survey results suggest that 98% of participants reported a positive experience with the software.
Levy et al (2011)	To examine learners' perceptions of sharing biometric data with university or third-party service providers.	Privacy, multimode biometrics, identity assurance.	Survey. 163 online students enrolled on an introductory IT course.	The findings suggest that online learners are more likely to enroll their biometric profile and use biometric authentication through their university than that of a third party service provider.

Lilley et al (2016)	To pilot test ProctorU remote invigilation service and examine learner attitudes towards its use.	Remote proctoring, identity assurance, video monitoring, usage restrictions, privacy.	Pilot study with 21 participants and survey completed by 9 participants at the University of Hertfordshire.	The results suggest that learner' overall attitude towards the use of remote proctoring were positive. Prior to the study, some participants expressed concerns about data protection and privacy.
McNabb and Maynard (2010)	To assess student acceptance of the use of Biometric signature for the purposes of authentication.	Identity assurance, biometric signature, mouse dynamics.	Pilot study.167 students completed the enrollment process and 73 students completed the survey at the University of Texas System.	The survey results suggest that enrollment and verification processes were easy to complete.
Migicovsky et al (2014)	To discuss the implication of wearable technology on academic misconduct; create a proof of concept system for cheating on a multiple choice test using a smart watch; propose methods of prevention.	Academic misconduct, authorship assurance, collusion, methods of prevention.	Proof of concept.	The results suggest that wearable technology may facilitate academic misconduct. Methods of prevention may mobile device policy restricting usage, question randomization.
Monaco et al (2013)	To examine performance of behavioral biometrics and stylometry.	Identity assurance, biometrics, authorship attribution.	Experiment. Replication of Stewart et al. (2011) study to evaluate the stylometry performance.	The keystroke system performance results on the student test data were 100% on the 6000 keystroke full-test and 99.96% on the 3000 keystroke half-test samples. Whereas stylometry system experiment of 30 book authors, yield 88.2% on 5000 words and 91.5% and 10000 words.
Mothukuri et al (2012)	To propose a model of secure remote assessment.	Identity assurance, proctoring, multi-factor authentication, feature restrictions, voiceprint, face recognition, collusion.	Theoretical framework.	Authors suggest a needs-based approach to assessment security. High-stakes examinations may require greater level of assurance than that of assignments.
Mott (2010)	To examine validity of a statistical algorithm that detects instances of collusion during concurrent assessment as well as examine the relationship between time allotted for the assessment, order of questions, ability to revisit questions once they had been answered.	Collusion detection, instructional design.	Experiment conducted during the 2009–2010 academic year at Purdue University with 53 Students enrolled in 2 aviation courses.	Findings suggest validity of the algorithm and positive correlation between frequency of cheating and time allotted for the assessment, as well as a negative correlation between randomization of test questions and frequency of cheating behavior.

Murphy and Holme (2015)	To discuss the implication of mobile phone imaging capabilities on academic misconduct during chemistry exams.	Academic misconduct, privacy, exam security, policy, methods of prevention.	Position article.	Mobile technology poses a concern. Solutions include: policy to restrict phone usage during exams. conduct regular online searches for compromised content, higher frequency of new tests, e-assessment.
O'Reilly and Creagh (2015)	To examine remote proctoring services, gaps in their quality control systems, and compare them to traditional invigilation methods.	Traditional proctoring, technology facilitated proctoring, automatic proctoring, quality control, identity assurance, authorship assurance.	Review.	Remote proctoring is a deterrent to cheating, although academic misconduct is still a possibility. IT Security issues may compromise delivery.
OReilly and Creagh (2016)	To discuss and categorize online proctoring services.	Traditional proctoring, technology facilitated proctoring, automatic proctoring, identity assurance, authorship assurance.	Review.	Proctoring services do not claim to be cheat proof. Automated detection of cheating requires manual verification. High stakes exams are better invigilated by human invigilators.
Pan et al (2004)	To describe a framework for conducting secure e-assessments within the existing network environment.	Secure assessment, automation, IP monitoring, data security.	Theoretical framework.	The framework is promising. Further research is needed to evaluate effectiveness of this approach.
Paullet et al (2014)	To examine identity assurance and academic integrity strategies employed in the distance education setting.	Identity assurance, authorship assurance, biometrics, keystroke dynamics, computer feature restrictions.	Survey of 457 students at 2 small mid-Atlantic Universities.	The results suggest that learner authentication is not consistently employed; the majority of students do not read academic integrity policy; proctoring may be employed by some universities.
Rabuzin et al (2006)	To review application of biometrics in distance learning environment.	Identity assurance, multimodal biometrics, privacy, data security.	Position article.	Identity assurance may be addressed through application of biometrics.
Ramim and Levy (2007)	To describe a framework for authentication during online assessment using a fingerprint scanner.	Identity assurance, unimodal biometrics, fingerprint, continuous authentication, collusion.	Theoretical framework.	Identity assurance may be addressed through application of biometrics. Further research is needed to evaluate effectiveness of the proposed model.
Ramu and Arivoli (2013)	To examine risks, benefits and limitations of the existing authentication approaches.	Identity assurance, anonymity, proxy test taking, keystroke dynamics, continuous authentication, IP monitoring, challenge questions.	Theoretical framework.	Authors propose to employ a combination of keystroke dynamics and challenge question authentication. Further research is needed to assess effectiveness of the proposed model.

Rodchua et al (2011)	To examine current biometric and proctoring technologies and propose a framework for secure e-assessment at the School of Technology, University of Central Missouri.	Identity assurance, proctoring, biometrics, video monitoring, feature restrictions, instructional design.	Theoretical framework.	Authors propose a model of remote proctoring that includes face recognition, feature restrictions, randomization of questions and timed exams.
Rosen and Carr (2013)	To report on the development of a new low-cost, autonomous desktop robot for proctoring examinations in online courses.	Identity assurance, asynchronous proctoring, ear recognition, biometrics, continuous authentication, data security.	Proof of concept, prototype.	Identity assurance and remote proctoring is possible. More research is needed to evaluate effectiveness of the proposed device.
Saunders et al (2008)	To examine perceptions of chairpersons of accounting departments at universities and colleges in the U.S on the issues of academic integrity and identity assurance.	Identity assurance, proctoring, video monitoring, instructor validation.	Survey of 56 chairpersons of accounting departments at universities and colleges across the U.S.	22.4% of the participants agreed that all examinations should be delivered online, whereas 38.8% disagreed with that notion. Approximately 25% of the participants agreed with a requirement to employ webcam monitoring for course registration, whereas 35% did not approve of this requirement.
Sheridan et al (2005)	To examine learners' perceptions of academic integrity issues and their experience with plagiarism detection tool TurnitIn.	Plagiarism detection, policy.	Case study. Survey of 110 third and fourth year students enrolled in the BPharm course at the School of Pharmacy, the University of Auckland.	The survey results suggest that students have a general understanding of the purpose of Turnitin. Access to the plagiarism checker enabled students to better understand academic integrity issues. 1/3 students perceived the use of the plagiarism checker as a form of distrust on the part of the school.
Shyles (2002)	To examine academic integrity issues and strategies.	Identity assurance, identity fraud, policies, proctoring, instructor validation, biometrics, authorship assurance, data security, privacy, accessibility.	Position article.	In the absence of measures that ensure academic integrity, credibility of distance education may be undermined.
Stewart et al (2011)	To examine performance of keystroke dynamics and stylometry biometric systems in distance learning environment.	Identity assurance, authorship attribution, biometrics, keystroke dynamics, stylometry.	Experiment. 30 students enrolled on a spreadsheet modeling course. The text lengths of the test answers ranged from 433 to 1831 words per test, with a mean of 966 words.	Asynchronous identification is possible and may satisfy the Higher Education Opportunities Act requirements.

| Ullah et al (2012) | To discuss performance and perception of usability of a framework for authenticating learners through a combination of password and challenge questions. | Authentication, identity assurance, authorship assurance, challenge questions, biometrics: fingerprint, voice, face recognition. | Experiment and Survey. 13 participants answered 39 challenge questions. | The results suggest a possible performance and usability issues. 62% of answers to challenge questions were successfully matched. Test interruption was perceived as inconvenience by a number of test takers. |

References

Aceves PA, Aceves RI (2009) Student identity and authentication in distance education: A primer for distance learning administrators. Continuing Higher Education Review 73:143–152

Afroz S, Brennan M, Greenstadt R (2012) Detecting hoaxes, frauds, and deception in writing style online. In: Security and Privacy (SP), 2012 IEEE Symposium on, IEEE, pp 461–475

Afroz S, Islam AC, Stolerman A, Greenstadt R, McCoy D (2014) Doppelgänger finder: Taking stylometry to the underground. In: 2014 IEEE Symposium on Security and Privacy (SP), IEEE, pp 212–226

Ali N, Hindi M, Yampolskiy RV (2011) Evaluation of authorship attribution software on a chat bot corpus. In: 2011 XXIII International Symposium on Information, Communication and Automation Technologies (ICAT), IEEE, pp 1–6

Allison B, Guthrie L (2008) Authorship attribution of e-mail: Comparing classifiers over a new corpus for evaluation. In: LREC

Alsallal M, Iqbal R, Amin S, James A (2013) Intrinsic plagiarism detection using latent semantic indexing and stylometry. In: 2013 Sixth International Conference on Developments in eSystems Engineering (DeSE), IEEE, pp 145–150

Amigud A (2013) Institutional level identity control strategies in the distance education environment: A survey of administrative staff. The International Review of Research in Open and Distributed Learning 14(5), DOI 10.19173/irrodl.v14i5.1541

Amigud A, Arnedo-Moreno J, Daradoumis T, Guerrero-Roldan AE (2016) A behavioral biometrics based and machine learning aided framework for academic integrity

in e-assessment. In: Intelligent Networking and Collaborative Systems (INCoS), 2016 International Conference on, IEEE, pp 255–262

Amigud A, Arnedo-Moreno J, Daradoumis T, Guerrero-Roldan AE (2017b) Open proctor: An academic integrity tool for the open learning environment. In: International Conference on Intelligent Networking and Collaborative Systems, Springer, pp 262–273

Amigud A, Arnedo-Moreno J, Daradoumis T, Guerrero-Roldan AE (2017a) Using learning analytics for preserving academic integrity. The International Review of Research in Open and Distributed Learning 18(5)

Antonia A, Craig H, Elliott J (2013) Language chunking, data sparseness, and the value of a long marker list: explorations with word n-grams and authorial attribution. Literary and Linguistic Computing 29(2):147–163

Apampa KM, Wills G, Argles D (2010) User security issues in summative e-assessment security. International Journal of Digital Society (IJDS) 1(2):1–13

Argamon S, Dhawle S, Koppel M, Pennebaker JW (2005) Lexical predictors of personality type. In: Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America

Argamon S, Koppel M, Pennebaker JW, Schler J (2009) Automatically profiling the author of an anonymous text. Communications of the ACM 52(2):119–123

Arun R, Suresh V, Madhavan CV (2009) Stopword graphs and authorship attribution in text corpora. In: 2009 International Conference on Semantic Computing (ICSC), IEEE, pp 192–196

Auernheimer B, Tsai MJ (2005) Biometric authentication for web-based couse examinations. In: Proceedings of the 38th Hawaii International Conference on System Sci-

ences, IEEE, p 294b

Ayodele TO (2010) Types of machine learning algorithms. INTECH Open Access Publisher

Baepler P, Murdoch CJ (2010) Academic analytics and data mining in higher education. International Journal for the Scholarship of Teaching and Learning 4(2):17

Bailie JL, Jortberg MA (2008) Distance learning student authentication: Verifying the identity of online students. A Collection of Paperson SelfStudy and Institutional Improvement p 65

Bailie JL, Jortberg MA (2009) Online learner authentication verifying the identity of online users. MERLOT Journal of Online Learning and Teaching 5(2)

Barnes C, Paris L (2013) An analysis of academic integrity techniques used in online courses at a southern university. In: Proceedings for the Northeast Region Decision Sciences Institute, pp T2D–17

Beasley EM (2014) Students reported for cheating explain what they think would have stopped them. Ethics & Behavior 24(3):229–252, DOI 10.1080/10508422.2013.845533

Bejarano AM, García LE, Zurek EE (2013) Detection of source code similitude in academic environments. Computer Applications in Engineering Education 23(1):13–22, DOI 10.1002/cae.21571

Bichsel J (2013) The state of e-learning in higher education: An eye toward growth and increased access. EDUCAUSE Center for Analysis and Research

Bird S (2006) Nltk: the natural language toolkit. In: Proceedings of the COLING/ACL on Interactive presentation sessions, Association for Computational Linguistics, pp 69–72

Black EW, Dawson K, Priem J (2008) Data for free: Using lms activity logs to measure community in online courses. The Internet and Higher Education 11(2):65–70

Blackmon SJ (2016) Through the MOOCing glass: Professors' perspectives on the future of MOOCs in higher education. New Directions for Institutional Research 2015(167):87–101, DOI 10.1002/ir.20156, URL

Bozkurt A, Akgun-Ozbek E, Yilmazel S, Erdogdu E, Ucar H, Guler E, Sezgin S, Karadeniz A, Sen-Ersoy N, Goksel-Canbek N, et al (2015) Trends in distance education research: A content analysis of journals 2009-2013. The International Review of Research in Open and Distributed Learning 16(1)

Bradbury R (1951) Fahrenheit 451: A Novel. Classic edition, Simon & Schuster

Brahnam S, Roberts JJ, Nanni L, Starr CL, Bailey SL (2015) Design of a bullying detection/alert system for school-wide intervention. In: International Conference on Human-Computer Interaction, Springer, pp 695–705

Brennan M, Afroz S, Greenstadt R (2012) Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. ACM Transactions on Information and System Security 15(3):1

Brocardo ML, Traore I, Saad S, Woungang I (2013) Authorship verification for short messages using stylometry. In: 2013 International Conference on Computer, Information and Telecommunication Systems (CITS), IEEE, pp 1–6

Brocardo ML, Traore I, Woungang I (2014) Toward a framework for continuous authentication using stylometry. In: 2014 IEEE 28th International Conference on Advanced Information Networking and Applications (AINA), IEEE, pp 106–115

Brocardo ML, Traore I, Woungang I (2015) Authorship verification of e-mail and tweet messages applied for continuous authentication. Journal of Computer and System Sci-

ences 81(8):1429–1440, DOI 10.1016/j.jcss.2014.12.019

Brooks DC, Thayer TLB (2016) Institutional analytics in higher education. Tech. rep.,
Research report. Louisville, CO: ECAR

Bruhn M, Gettes M, West A (2003) Identity and access management and security in
higher education. Educause Quarterly 26(4):12–17

Buzdar MA, Ali A, Tariq RUH (2016) Emotional intelligence as a determinant of readi-
ness for online learning. The International Review of Research in Open and Distributed
Learning 17(1), DOI 10.19173/irrodl.v17i1.2149

Campbell JP, Oblinger DG, et al (2007) Academic analytics. EDUCAUSE review
42(4):40–57

Canales O, Monaco V, Murphy T, Zych E, Stewart J, Castro C, Sotoye O, Torres L,
Truley G (2011) A stylometry system for authenticating students taking online tests. P
of Student-Faculty Research Day, Ed, CSIS Pace University

Carr M, Verner J (1997) Prototyping and software development approaches. Depart-
ment of Information Systems, City University of Hong Kong, Hong Kong URL
http://www.uwex.edu/disted/conference/Resource_library/proceedings

Case R, Cabalka P (2009) Remote proctoring: results of a pi-
lot program at western governors university. 09_19933.pdf, URL
http://www.uwex.edu/disted/conference/Resource_library/
proceedings/09_19933.pdf

Chaski CE (2005) Who's at the keyboard? authorship attribution in digital evidence in-
vestigations. International Journal of Digital Evidence 4(1):1–13

Chaski CE (2012) Best practices and admissibility of forensic author identification. JL &
Pol'y 21:333

Chatti MA, Dyckhoff AL, Schroeder U, Thüs H (2012) A reference model for learning analytics. International Journal of Technology Enhanced Learning 4(5-6):318–331

Cheng N, Chandramouli R, Subbalakshmi K (2011) Author gender identification from text. Digital Investigation 8(1):78–88

Chiesl N (2007) Pragmatic methods to reduce dishonesty in web-based courses. Quarterly Review of Distance Education 8:203–211

Chiranji N, Deepthi C, Shekhar T (2011) A novel approach to enhance security for online exams. International Journal on Computer Science and Technology 2(3):85–90

Clarke NL, Dowland P, Furnell SM (2013) e-invigilator: A biometric-based supervision system for e-assessments. In: Information Society (i-Society), 2013 International Conference on, IEEE, pp 238–242

Clow D (2012) The learning analytics cycle: closing the loop effectively. In: Proceedings of the 2nd international conference on learning analytics and knowledge, ACM, pp 134–138

Cluskey Jr G, Ehlen CR, Raiborn MH (2011) Thwarting online exam cheating without proctor supervision. Journal of Academic and Business Ethics 4:1

Collins C, Carpendale S, Penn G (2009) Docuburst: Visualizing document content using language structure. In: Computer graphics forum, Wiley Online Library, vol 28, pp 1039–1046

Colwell J, Jenks C (2005) Student ethics in online courses. In: Proceedings Frontiers in Education 35th Annual Conference, IEEE, DOI 10.1109/fie.2005.1611897

Cooper HM (1984) The integrative research review: A systematic approach. Sage

Cortes C, Kuznetsov V, Mohri M (2014) Ensemble methods for structured prediction. In: International Conference on Machine Learning, pp 1134–1142

Crawford G, Rudy JA (2003) Fourth annual EDUCAUSE survey identifies current IT issues. Educause Quarterly 26(2):12–27

Culwin F (2008) A longitudinal study of nonoriginal content in final-year computing undergraduate projects. IEEE Transactions on Education 51(2):189–194, DOI 10.1109/te.2007.910350

Culy C, Lyding V (2010) Double tree: an advanced kwic visualization for expert users. In: 2010 14th International Conference Information Visualisation, IEEE, pp 98–103

Ding SH, Fung B, Debbabi M (2015) A visualizable evidence-driven approach for authorship attribution. ACM Transactions on Information and System Security (TISSEC) 17(3):12

Domingos P (2012) A few useful things to know about machine learning. Communications of the ACM 55(10):78–87

Dutt A, Ismail MA, Herawan T (2017) A systematic review on educational data mining. IEEE Access

El Bouanani SEM, Kassou I (2014) Authorship analysis studies: A survey. International Journal of Computer Applications 86(12):22–29, DOI 10.5120/15038-3384

Elbadrawy A, Polyzou A, Ren Z, Sweeney M, Karypis G, Rangwala H (2016) Predicting student performance using personalized analytics. Computer 49(4):61–69

Faith R (2016) How does the use of mobile phones by 16-24 year old socially excluded women affect their capabilities? PhD thesis, The Open University

Farias DIH, Guzman-Cabrera R, Reyes A, Rocha MA (2013) Semantic-based features for author profiling identification: First insights notebook for pan at clef 2013. In: Proceedings of CLEF

Fawcett T (2006) An introduction to roc analysis. Pattern recognition letters 27(8):861–874

Fendler RJ, Godbey JM (2015) Cheaters should never win: Eliminating the benefits of cheating. Journal of Academic Ethics 14(1):71–85, DOI 10.1007/s10805-015-9240-8

Fiedler R, Kaner C (2010) Plagiarism-detection services: How well do they actually perform? IEEE Technology and Society Magazine 29(4):37–43, DOI 10.1109/mts.2010.939225

Fifield D, Follan T, Lunde E (2015) Unsupervised authorship attribution. arXiv preprint arXiv:150307613

Forsyth RS, Holmes DI (1996) Feature-finding for text classification. Literary and Linguistic Computing 11(4):163–174

Foster D, Matton N, Walker P (2009) Using multiple online security measures to deliver secure course exams to distance education students. Final_Paper_101Walker.pdf, URL http://www.ou.nl/Docs/Campagnes/ICDE2009/Papers/Final_Pa per_101Walker.pdf

Frank AJ (2010) Dependable distributed testing: Can the online proctor be reliably computerized? In: e-Business (ICE-B), Proceedings of the 2010 International Conference on, IEEE, pp 1–10

Fridman L, Stolerman A, Acharya S, Brennan P, Juola P, Greenstadt R, Kam M (2015) Multi-modal decision fusion for continuous authentication. Computers & Electrical Engineering 41:142–156

Gao Q (2010) Online teaching: Do you know who is taking the final exam? In: American Society for Engineering Education (ASEE), Mid-Atlantic Conference, Villanova University

Gao Q (2012) Using ip addresses as assisting tools to identify collusions. International Journal of Business, Humanities and Technology 2(1):70–75

Gašević D, Dawson S, Rogers T, Gasevic D (2016) Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. The Internet and Higher Education 28:68–84

Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Machine learning 63(1):3–42

Gibbs G, Simpson C (2004) Conditions under which assessment supports studentsâ™ learning. Learning and teaching in higher education 1(1):3–31

Gregory J, Strukov A (2002) Ensuring academic integrity in the age of the internet: evaluating a web-based analytic tool. In: International Conference on Computers in Education, 2002. Proceedings., IEEE Comput. Soc, DOI 10.1109/cie.2002.1186250

Greller W, Drachsler H (2012) Translating learning into numbers: A generic framework for learning analytics. Educational technology & society 15(3):42–57

Gudivada V, Irfan M, Fathi E, Rao D (2016) Cognitive analytics: Going beyond big data analytics and machine learning. Handbook of Statistics 35:169–205

Gullifer J, Tyson G (2014) Who has read the policy on plagiarism? unpacking students' understanding of plagiarism. Studies in Higher Education 39(7):1202–1218, DOI 10.1080/03075079.2013.777412

Heckler NC, Rice M, Hobson Bryan C (2013) Turnitin systems: A deterrent to plagiarism in college classrooms. Journal of Research on Technology in Education 45(3):229–248

Herschel G, Linden A, Kart L (2015) Magic quadrant for advanced analytics platforms. Gartner Report G 270612

Holmes DI (1997) Stylometry: its origins, development and aspirations. In: Joint international conference of the association for computers and the humanities and the association for literary and linguistic computing, pp 98–103

Holmes DI (1998) The evolution of stylometry in humanities scholarship. Literary and linguistic computing 13(3):111–117

Hoshiar M, Dunlap J, Li J, Friedel JN (2014) Examining the effectiveness of student authentication and authenticity in online learning at community colleges. Community College Journal of Research and Practice 38(4):337–345, DOI 10.1080/10668926.2012.755649

Howedi F, Mohd M (2014) Text classification for authorship attribution using naive bayes classifier with limited training data. Computer Engineering and Intelligent Systems 5(4):48–56

Iqbal F, Binsalleeh H, Fung BC, Debbabi M (2013) A unified data mining solution for authorship analysis in anonymous textual communications. Information Sciences 231:98–112

Irfan A, Nomura S, Ouzzane K, Fukumura Y (2009) Face-based access control and invigilation tool for e-learning systems. In: 2009 International Conference on Biometrics and Kansei Engineering, IEEE, DOI 10.1109/icbake.2009.43

ITU (2012) Entity authentication assurance framework.recommendation itu-t x.1254. URL http://www.itu.int/rec/T- REC- X.1254- 201209- I

Jain A, Ross A, Prabhakar S (2004) An introduction to biometric recognition. IEEE Transactions on Circuits and Systems for Video Technology 14(1):4–20, DOI 10.1109/tcsvt.2003.818349

Jiao P (2011) How to solve the security and privacy problems within e-learning. In: 2011 IEEE International Symposium on IT in Medicine and Education, IEEE, DOI 10.1109/itime.2011.6130785

Jockers ML, Witten DM (2010) A comparative study of machine learning methods for authorship attribution. Literary and Linguistic Computing 25(2):215–223, DOI 10.1093/llc/fqq001

Johnson A, Wright D (2014) Identifying idiolect in (forensic) authorship attribution: an n-gram textbite approach. Language and Law Linguagem e Direito 1(1):37–69

Joorabchi A, English M, Mahdi AE (2016) Text mining stackoverflow: an insight into challenges and subject-related difficulties faced by computer science learners. Journal of Enterprise Information Management 29(2):255–275

Jortberg M (2010) Experiences verifying the identity of distance learning students. Third annual report on identity in distance learning URL http://www.uwex.edu/disted/conference/Resource_library/proceedings

Juola P (2004) Ad-hoc authorship attribution competition. online http://www.mathcs.duq.edu/j̃uola/authorship_contest.html

Juola P (2006) Authorship attribution. Foundations and Trends in information Retrieval 1(3):233–334

Juola P, Baayen RH (2005) A controlled-corpus experiment in authorship identification by cross-entropy. Literary and Linguistic Computing 20(Suppl 1):59–67, DOI 10.1093/llc/fqi024

Kang BH, Kim H (2015) Proposal: A design of e-learning user authentication system. International Journal of Security and Its Applications 9(1):45–50, DOI 10.14257/ijsia.2015.9.1.05

Kermek D, Novak M (2016) Process model improvement for source code plagiarism detection in student programming assignments. INFORMATICS IN EDUCATION 15(1):103–126, DOI 10.15388/infedu.2016.06

Khalil M, Ebner M (2015) Learning analytics: principles and constraints. In: Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, pp 1326–1336

Kitchens MB (2014) Word clouds: An informal assessment of student learning. College Teaching 62(3):113–114

Koppel M, Schler J, Argamon S, Winter Y (2012) The "fundamental problem" of authorship attribution. English Studies 93(3):284–291

Lang J (2013) Cheating Lessons: learning from academic dishonesty. Harvard University Press

Larose DT (2005) Discovering knowledge in data : an introduction to data mining. John Wiley Sons, Inc., Hoboken, New Jersey

Levey S, Maynard J (2011) Identity proofing for online student id verification: Report of pilot with houston community college. BioSig-Study.pdf, URL http://de.hccs.edu/media/houston-community-college/distance-education/about-us-pdf/BioSig-Study.pdf

Levy Y, Ramim MM, Furnell SM, Clarke NL (2011) Comparing intentions to use university-provided vs vendor-provided multibiometric authentication in online exams. Campus-Wide Information Systems 28(2):102–113

Lilley M, Meere J, Barker T (2016) Remote live invigilation: A pilot study. Journal of Interactive Media in Education 2016(1), DOI 10.5334/jime.408

Lin Y, Michel JB, Aiden EL, Orwant J, Brockman W, Petrov S (2012) Syntactic anno-
tations for the google books ngram corpus. In: Proceedings of the ACL 2012 system
demonstrations, Association for Computational Linguistics, pp 169–174

Liu S, Liu Z, Sun J, Liu L (2011) Application of synergetic neural network in online
writeprint identification. International Journal of Digital Content Technology and its
Applications 5(3):126–135

Lopez D, Herrero JR, Pajuelo A, Duran A (2007) A proposal for continuous assess-
ment at low cost. In: 2007 37th Annual Frontiers In Education Conference - Global
Engineering: Knowledge Without Borders, Opportunities Without Passports, IEEE,
DOI 10.1109/fie.2007.4417840

Louppe G, Wehenkel L, Sutera A, Geurts P (2013) Understanding variable importances
in forests of randomized trees. In: Advances in neural information processing systems,
pp 431–439

Love H (2002) Attributing authorship: an introduction. Cambridge University Press

Luyckx K, Daelemans W (2005) Shallow text analysis and machine learning for author-
ship attribtion. LOT Occasional Series 4:149–160

Luyckx K, Daelemans W (2008a) Authorship attribution and verification with many au-
thors and limited data. In: Proceedings of the 22nd International Conference on Com-
putational Linguistics-Volume 1, Association for Computational Linguistics, pp 513–
520

Luyckx K, Daelemans W (2008b) Personae: a corpus for author and personality predic-
tion from text. In: LREC

Magnusson C, Vanharanta H (2003) Visualizing sequences of texts using collocational
networks. In: International Workshop on Machine Learning and Data Mining in Pattern

Recognition, Springer, pp 276–283

Marcus MP, Marcinkiewicz MA, Santorini B (1993) Building a large annotated corpus of english: The penn treebank. Computational linguistics 19(2):313–330

McGill S (2012) Plagiarism in Latin literature. Cambridge University Press

McNabb L, Maynard J (2010) Software only biometrics to authenticate student id report of pilot with the university of texas system telecampus. BSI_and_UTTC_case_study.pdf

Michaelson S, Morton A (1972) The new stylometry: a one-word test of authorship for greek writers. The Classical Quarterly (New Series) 22(1):89–102

Michel JB, Shen YK, Aiden AP, Veres A, Gray MK, Pickett JP, Hoiberg D, Clancy D, Norvig P, Orwant J, et al (2011) Quantitative analysis of culture using millions of digitized books. science 331(6014):176–182

Migicovsky A, Durumeric Z, Ringenberg J, Halderman JA (2014) Outsmarting proctors with smartwatches: A case study on wearable computing security. In: Financial Cryptography and Data Security, Springer Berlin Heidelberg, pp 89–96, DOI 10.1007/978-3-662-45472-5$_7$

Miller GA (1995) Wordnet: a lexical database for english. Communications of the ACM 38(11):39–41

Monaco JV, Stewart JC, Cha SH, Tappert CC (2013) Behavioral biometric verification of student identity in online course assessment and authentication of authors in literary works. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), IEEE, pp 1–8

Moore H (2014) Cheating students punished by the 1000s, but many more go undetected

Moore M, Kearsly G (2005) Distance education: A systems view of online learning (2nd ed.). Belmont, CA :Wadsworth, Cengage Learning

Moore MG, Kearsley G (1996) Distance education: A systems view. Boston, MA: Wadsworth

Mothukuri U, Jain S, Muralidharan V (2012) Invigilated online assessment: Various ways to minimize unauthorized help. In: E-Learning, E-Management and E-Services (IS3e), 2012 IEEE Symposium on, pp 1–4

Mott JH (2010) The detection and minimization of cheating during concurrent online assessments using statistical methods. Collegiate Aviation Review 28(2):32

Murphy KL, Holme TA (2015) What might cell phone-based cheating on tests mean for chemistry education? Journal of Chemical Education 92(9):1431–1432, DOI 10.1021/acs.jchemed.5b00637

Nirkhi SM, Dharaskar R, Thakare V (2015) Authorship identification using generalized features and analysis of computational method. Transactions on Machine Learning and Artificial Intelligence 3(2):41

Nitta T, Masui F, Ptaszynski M, Kimura Y, Rzepka R, Araki K (2013) Detecting cyberbullying entries on informal school websites based on category relevance maximization. In: IJCNLP, pp 579–586

Noecker J, Ryan M, Juola P (2013) Psychological profiling through textual analysis. Literary and Linguistic Computing 28(3):382–387

O'Reilly G, Creagh J (2015) Does the shift to cloud delivery of courses compromise quality control. In: Higher Education in Transformation Conference (HEIT), Dublin Institute of Technology, pp 417–424

OReilly G, Creagh J (2016) A categorization of online proctoring. In: Global Learn, vol 2016, pp 542–552

Pan CC, Yang KH, Lee TL (2004) Secure online examination architecture based on distributed firewall. In: IEEE International Conference on e-Technology, e-Commerce and e-Service, 2004. EEE 04. 2004, IEEE, DOI 10.1109/eee.2004.1287359

Pateriya PK, Raj G, et al (2014) A pragmatic validation of stylometric techniques using bpa. In: Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference-, IEEE, pp 124–131

Paullet K, Douglas DM, Chawdhry A (2014) Verifying user identities in distance learning courses: Do we know who is sitting and submitting behind the screen? Issues in Information Systems 15(1):370–379

Pavelec D, Oliveira LS, Justino E, Neto FN, Batista LV (2009) Compression and stylometry for author identification. In: 2009 International Joint Conference on Neural Networks, IEEE, pp 2445–2450

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12:2825–2830

Rabuzin K, Baca M, Sajko M (2006) E-learning: Biometrics as a security factor. In: International Multi-Conference on Computing in the Global Information Technology - (ICCGI06), IEEE, DOI 10.1109/iccgi.2006.1690338

Ramim MM, Levy Y (2007) Towards a framework of biometrics exam authentication in e-learning environments. In: Proceeding of the Information Resources Management Association International Conference (IRMA) 2007, pp 539–549

Ramos J (2003) Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning

Ramu T, Arivoli T (2013) A framework of secure biometric based online exam authentication: an alternative to traditional exam. Int J Sci Eng Res 4(11):52–60

Raschka S (2015) Python Machine Learning. Packt Publishing Ltd

Riemenschneider CK, Manly TS, Leonard LN (2016) Using giving voice to values to improve student academic integrity in information technology contexts. Journal of Information Systems Education 27(3):183

Rodchua S, Yiadom-Boakye G, Woolsey R (2011) Student verification system for online assessments: Bolstering quality and integrity of distance learning. Journal of Industrial Technology 27(3):1–8

Roffo G, Cristani M, Bazzani L, Minh HQ, Murino V (2013) Trusting skype: Learning the way people chat for fast user recognition and verification. In: 2013 IEEE International Conference on Computer Vision Workshops (ICCVW), IEEE, pp 748–754

Rosen WA, Carr ME (2013) An autonomous articulating desktop robot for proctoring remote online examinations. In: Frontiers in Education Conference, 2013 IEEE, IEEE, pp 1935–1939

Rudman J (1997) The state of authorship attribution studies: Some problems and solutions. Computers and the Humanities 31(4):351–365

Rusu D, Fortuna B, Mladenic D, Grobelnik M, Sipoš R (2009) Document visualization based on semantic graphs. In: 2009 13th International Conference Information Visualisation, IEEE, pp 292–297

Saevanee H (2015) Continuous user authentication using multi-modal biometrics. Computers & Security

Salmon G, Gregory J, Dona KL, Ross B (2015) Experiential online development for educators: The example of the carpe diem MOOC. British Journal of Educational Technology

46(3):542–556, DOI 10.1111/bjet.12256

Sapkota U, Bethard S, Montes-y Gómez M, Solorio T (2015) Not all character n-grams are created equal: A study in authorship attribution. In: Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pp 93–102

Saunders G, Wenzel L, Stivason CT (2008) Internet courses: Who is doing the work? Journal of College Teaching & Learning (TLC) 5(6), DOI 10.19030/tlc.v5i6.1251

Schmid MR, Iqbal F, Fung BC (2015) E-mail authorship attribution using customized associative classification. Digital Investigation 14:S116–S126, DOI 10.1016/j.diin.2015.05.012, URL http://dx.doi.org/10.1016/j.diin.2015.05.012

Seroussi Y, Zukerman I, Bohnert F (2011) Authorship attribution with latent dirichlet allocation. In: Proceedings of the fifteenth conference on computational natural language learning, Association for Computational Linguistics, pp 181–189

Sheridan J, Alany R, Brake DJ (2005) Pharmacy students' views and experiences of turnitin®â"an online tool for detecting academic dishonesty. Pharmacy Education 5(3-4):241–250

Shyles L (2002) Authenticating, identifying, and monitoring learners in the virtual classroom: Academic integrity in distance learning. Paper presented at the meeting of the National Communication Association, New Orleans, LA

Sidorov G, Velasquez F, Stamatatos E, Gelbukh A, Chanona-Hernández L (2014) Syntactic n-grams as machine learning features for natural language processing. Expert Systems with Applications 41(3):853–860

Siemens G (2012) Learning analytics: envisioning a research discipline and a domain of practice. In: Proceedings of the 2nd international conference on learning analytics and knowledge, ACM, pp 4–8

Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Information Processing & Management 45(4):427–437

Solorio T, Pillay S, Raghavan S, Montes-y Gómez M (2011) Modality specific meta features for authorship attribution in web forum posts. In: IJCNLP, pp 156–164

Stamatatos E (2007) Author identification using imbalanced and limited training texts. In: The 18th International Workshop on Database and Expert Systems Applications (DEXA'07), IEEE, pp 237–241

Stamatatos E (2009) A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology 60(3):538–556

Stamatatos E, Daelemans W, Verhoeven B, Potthast M, Stein B, Juola P, Sanchez-Perez MA, Barrón-Cedeño A (2014) Overview of the author identification task at pan 2014. CLEF 2014 Evaluation Labs and Workshop pp 877–897

Stamatatos E, Potthast M, Rangel F, Rosso P, Stein B (2015) Overview of the PAN/CLEF 2015 evaluation lab. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer, pp 518–538, DOI 10.1007/978-3-319-24027-5$_4$9

Stein B, Rosso P, Stamatatos E, Koppel M, Agirre E (2009) 3rd pan workshop on uncovering plagiarism, authorship and social software misuse. In: 25th Annual Conference of the Spanish Society for Natural Language Processing (SEPLN), pp 1–77

Stein B, Lipka N, Prettenhofer P (2011) Intrinsic plagiarism analysis. Language Resources and Evaluation 45(1):63–82

Stewart JC, Monaco JV, Cha SH, Tappert CC (2011) An investigation of keystroke and stylometry traits for authenticating online test takers. In: Biometrics (IJCB), 2011 International Joint Conference on, IEEE, pp 1–7

Stuart LM, Tazhibayeva S, Wagoner AR, Taylor JM (2013) On identifying authors with style. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, pp 3048–3053

Tschuggnall M, Gunter S (2014) What grammar tells about gender and age of authors. In: Proceedings of the Fourth International Conference on Advances in Information Mining and Management, IARIA, pp 30–35

Ullah A, Xiao H, Lilley M, Barker T (2012) Using challenge questions for student authentication in online examination. International Journal for Infonomics (IJI) 5(3/4):9

Unar J, Seng WC, Abbasi A (2014) A review of biometric technology along with trends and prospects. Pattern Recognition 47(8):2673–2688, DOI 10.1016/j.patcog.2014.01.016, URL

Vuillemot R, Clement T, Plaisant C, Kumar A (2009) What's being said near "martha"? exploring name entities in literary text collections. In: Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on, IEEE, pp 107–114

Wan CH, Lee LH, Rajkumar R, Isa D (2012) A hybrid text classification approach with low dependency on parameter by integrating k-nearest neighbor and support vector machine. Expert Systems with Applications 39(15):11,880–11,888

Wattenberg M, Viégas FB (2008) The word tree, an interactive visual concordance. IEEE transactions on visualization and computer graphics 14(6):1221–1228

West DM (2012) Big data for education: Data mining, data analytics, and web dashboards. Governance Studies at Brookings 4:1–0

Zhao Y, Zobel J (2007a) Entropy-based authorship search in large document collections. Springer

Zhao Y, Zobel J (2007b) Searching with style: authorship attribution in classic literature. In: Proceedings of the thirtieth Australasian conference on Computer science-Volume 62, Australian Computer Society, Inc., pp 59–68

Zhong N, Li Y, Wu ST (2012) Effective pattern discovery for text mining. IEEE transactions on knowledge and data engineering 24(1):30–44