reconnect
*Reconnecting relationships through technology*

# The Impact of Cognitive Load on Students' Academic Writing: An Authorship Verification Investigation

**Eduardo Araujo Oliveira, Paula De Barba**
The University of Melbourne

Automatic authorship verification is known to be a challenging machine learning task. In this paper, we examine the efficacy of an enhanced common n-gram profile-based approach to assist educational institutions to validate students' essays and assignments through their writing styles. We investigated the impact that essays with different cognitive load requirements have in students' writing styles, which may or may not impact authorship verification methods. A total of 46 undergraduate students completed six essays in a laboratory study. Although results showed small and mixed effects of the tasks differing in cognitive load on the different writing product metrics, students' essays and assignments texts contained features that remained stable across essays requiring different levels of cognitive load. These results suggest that our approach could be successfully used in authorship verification, potentially helping to address issues related to academic integrity in higher education settings.

Keywords: academic integrity, authorship verification, writing analytics, learning analytics, stylometry, cognitive load.

## Introduction

Academic integrity is a growing issue facing higher education institutions, with increasing numbers of reported academic fraud worldwide. This issue is related, at least to some extent, to the quick growth of universities and higher education systems (Macfarlane et al., 2014). Although it is unclear what is the best course of action on how to deal with academic integrity, universities have high stakes on guaranteeing that their graduates will uphold their institutions' reputation once in the workforce (Awdry et al., 2021). Automated authorship verification is a technology that universities could use to monitor students' academic integrity at scale.

Authorship verification (AV) in higher education has potential to be applied to essays, a widely used form of assessment. This technology relies on applying algorithms to detect whether students are the author of submitted essays, based on their writing styles (i.e., stylometry). This is a useful technology for contract cheating, which is when students outsource essay writing to either companies ("essay mills") or friends and family. However, there are some challenges in the implementation of stylometry in higher education. Even though previous research has found students' writing style varied across essay tasks with different levels of difficulty (i.e., cognitive load) (Oliveira et al., 2020), it is unknown whether these variations would impact authorship verification. Cognitive load reflects the notion that a student's ability to perform a task depends on the cognitive demands of the task, and the student's working memory capacity available for task processing (Sweller, 1988). If the cognitive demands required for a given task exceed students' available working memory capacity, students' ability to perform the task will be affected. Students may take longer to process information, use strategies that require less cognitive load, or make more errors (Beilock & DeCaro, 2007; Parkman & Groen, 1971). Writing is a complex cognitive task, requiring coordination of long-term knowledge, language skills, motor skills, and working memory. This means that an authorship verification method could be unable to identify the same author across essays with different levels of difficulty.

In this context, this project aims to evaluate potential automated authorship identification or attribution technology to assist educational institutions to validate students' essays and assignments through their writing styles. As such, this paper extends research initiated by Potha and Stamatatos (2014) and Oliveira and colleagues (2020), evaluating and discussing the effectiveness and accuracy of an enhanced Common-N-Gram (CNG) profile-based approach combined with an investigation on the impact of essays with different cognitive load requirements.

## Background literature

### Essay writing and cognitive load in higher education

Essay writing is a widely used form of assessment in higher education and it can be used to assess different learning objectives (Brizan et al., 2015). The Bloom taxonomy proposes six educational objectives: (1) remember, e.g., retrieval, (2) understand, e.g., interpret and explain, (3) apply, e.g., execute and implement, (4) analyse, e.g., organise and attribute, (5) evaluate, e.g., critique and make judgements, (6) create, e.g., generate and plan (Anderson & Krathwohl, 2001). These categories are thought to increasingly demand higher cognitive load from students (Brizan et al., 2015). That is, essay requiring students to remember or explain something are thought to demand students' working memory to hold less information at one time than essays requiring them to analyse or create something. If the cognitive demands required for a given task exceed students' available working memory capacity, students' ability to perform the task will be affected. Students may take longer to process information, use strategies that require less cognitive load, or make more errors.

Previous research has found that such differences in cognitive load demands can be detected in essay writing using writing analytics (Oliveira et al., 2020). In the current study, we focus on the writing product or final essays and assignment texts submitted by students. Stylometry is used to analyse static completed texts (i.e., product). Stylometry is based on the linguistic style of the text produced by the author (Calix et al., 2008). The style of a completed text can be characterised by measuring a vast array of stylistic features, that includes lexical (e.g., word, sentence or character-based statistic variation such as vocabulary richness and word-length distributions), syntactic (e.g., function words, punctuation and part-of-speech), structural (e.g., text organisation and layout, fonts, sizes and colours), content-specific (e.g., word n-grams), and idiosyncratic style markers (e.g., misspellings, grammatical mistakes and other us age anomalies) (Abbasi & Chen, 2008; Holmes & Kardos, 2003). Stylometry is often used for authorship identification.

### Authorship identification

Automated authorship identification or attribution is the problem concerned in identifying the true author of an anonymous document given samples of undisputed documents from a set of candidate authors (Keselj et al., 2003). The identification of authors is inferred from modeling of writing styles (Mosteller & Wallace, 1963; Potthast et al., 2016; Potha & Stamatatos, 2014) and its attribution is often examined in the relevant literature in three main forms: (i) open-set attribution, when the candidate authors may not contain the true author of some of the questioned documents (Potha & Stamatatos, 2014), (ii) authorship verification, when given examples of the writing of a single author, the aim is to determine if new texts were or were not written by the same author (Koppel & Schler, 2004; Potha & Stamatatos, 2014) and, (iii) closed-set attribution, when the candidate authors include the true authors of questioned documents (Potha & Stamatatos, 2014; Koppel & Winter, 2014). According to Potha and Stamatatos (2014), all authorship attribution cases can be transformed to different sets of authorship verification problems. As a categorisation problem, authorship verification is more complex than the other authorship attribution forms because a single author may intentionally vary his or her style from text to text for many reasons or may unconsciously drift stylistically over time (Koppel & Schler, 2004).

The use of stylometry for authorship identification assumes that an author's writing style is consistent and recognisable (Laramee, 2018). Stylistic features are the attributes or writing-style markers that are the most effective discriminators of authorship. Over 1000 different style markers have been used in previous research on stylistic analysis, with no consensus on the best set (Rudman, 1997).

### Authorship verification and essays writing with different cognitive loads

Attempts to solve authorship attribution problems follow either the instance-based or the profile-based paradigm. The instance-based paradigm treats all available samples by one author separately; in this paradigm each text sample has its own representation. On the other hand, the profile-based paradigm treats all available text samples by one candidate author cumulatively. Text samples are concatenated into a single, often large representative document and then the profile of the author is extracted from that document (Potha & Stamatatos, 2014). Another profile is produced from the questioned document and the two profiles are compared using a dissimilarity function. Due to constant changes and improvements on students' vocabularies among higher education courses, the profile-based paradigm will be combined and investigated together with the CNG method in this study. We believe this paradigm can help us to establish and maintain students' profiles across several years while providing more flexibility and higher accuracy in authorship verification.

In a previous study, Oliveira and colleagues (2020) focused on writing analytics, they asked students to complete four activities distributed over a period of 90 minutes. To account for possible effects of question ordering, two setups were used: one setup with increasing cognitive load, from low (1) to high (6) and one setup with decreasing cognitive load, from high (6) to low (1). The first 29 participants completed Setup 1, while the following 17 participants completed Setup 2. In this study, the authors used seven metrics (percentage of sentence linking connectives, semantic similarity, mean length of T-unit, clause density, mean word frequency, percentage of long words and percentage of misspelled words) across four dimensions to analyse the writing outcome. The results showed only small and mixed effects of the tasks differing in cognitive load on the different writing product metrics. Students writing products remained stable and consistent across different cognitive loads.

## Current study

In the current study we examine whether an AV algorithm would be able to identify the same author across essays with different cognitive load requirements in educational settings. That is, we evaluate and discuss the CNG profile-based paradigm efficiency and accuracy in supporting authorship verification of essays and assignments with different cognitive loads in higher education.

## Method

Following the proposed approach by Castro and colleagues (Castro et al., 2015) related to method verification in text analyses (PAN dataset), our method included data collection, data pre-processing, authorship verification method analysis (Study 1) and main data analyses (Studies 2 and 3). These steps are presented in Figure 1.

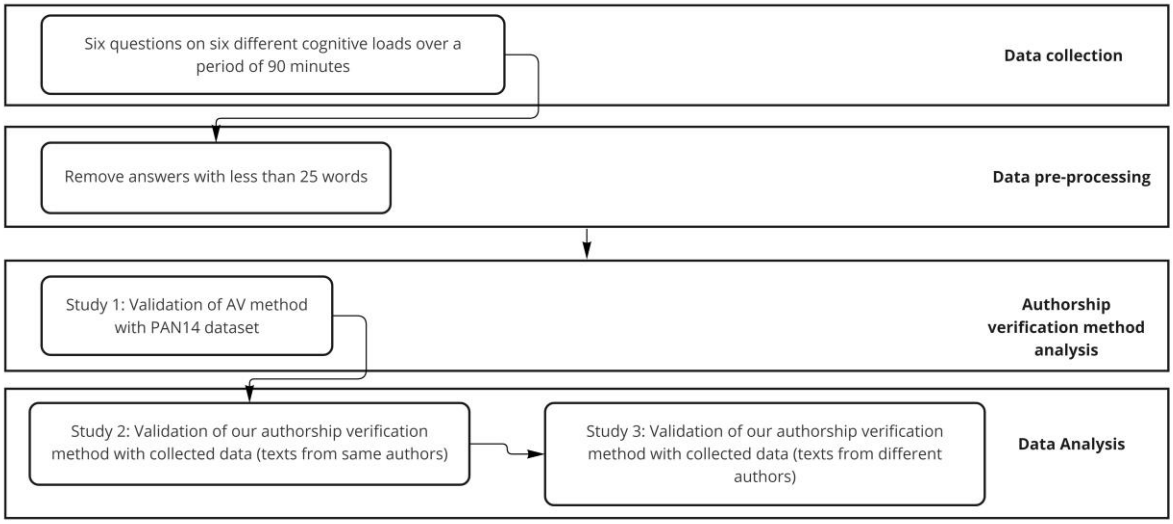


**Figure 1: Research procedure**

### Participants

The study was conducted at The University of Melbourne from 2017 to 2019. Participants were recruited via posters across the campus and provided informed consent (Ethics approval #1748727.1). The sample included a total of 46 students from four main disciplines: Engineering (24%, n=11), Commerce (24%, n=11), Arts (19.5%, n=9), Science (13%, n=6) and other (19.5%, n=9). Most participants were undergraduate students (70%, n=32), with 24 males (52%) and 22 females (48%). More than half of the participants were from a non-English speaking background (76%, n=35), and most participants were right-handed (96%, n=44).

### Data collection

In a computer laboratory, participants were asked to complete four activities using an Apple desktop computer and a QWERTY keyboard. The four activities were distributed over a period of 90 minutes (Figure 1). To account for possible effects of question ordering, two setups were used: one setup with increasing cognitive load, from low (1) to high (6) and one setup with decreasing cognitive load, from high (6) to low (1), as shown

in Figure 1 The first 29 participants completed Setup 1, while the following 17 participants completed Setup 2. In the Creative Work 1 activity participants had 20 minutes to answer four open-ended questions requiring low to medium cognitive load (Q1, Q2, Q3, Q4; see Table 1). In the Creative Work 2 activity participants had 30 minutes to answer two open-ended questions requiring medium to high cognitive load (Q5, Q6; see Table 1). For the questions that required medium to high cognitive load, participants could consult two hardcopy supporting texts on the topic of university life. Participants then had a 10-minute break, where some snacks were provided. In the Review activity, participants had 10 minutes to review, edit and improve their answers from the Creative Work 2 activity (Q5a, Q6a; see Table 1). In Transcription activity participants were asked to transcribe one of the texts that was used as a support material during 'Creative Work 2' for 10 minutes (Q7).

**Table 1: List of questions and respective level of cognitive load.**

| ID | CL | Question |
|---|---|---|
| Q1 | 1 | What made you decide to join this university? |
| Q2 | 2 | What would you say has been the best class you have taken at this university and what did you enjoy about that class? |
| Q3 | 3 | You are asked to complete a group assignment. It is important all students in the group contribute equally to the project. Come up with a plan for completing the group assignment, from research to class presentation. |
| Q4 | 4 | Describe the similarities and differences between preparing a written assignment and preparing for a final exam. |
| Q5 | 5 | A fellow university student spends a significant amount of their time worrying about their ability to complete their academic work, and becomes very concerned when they do not meet their grade expectations. In addition, they are concerned about financial pressures such as rent and textbook costs. Considering the texts you have received and the situation presented above, please answer the following question: Do you think the university should support this student improve their wellbeing? Why or why not? |
| Q6 | 6 | [Using the scenario from Q5] Describe what advice you would provide to the student to help improve their wellbeing. What steps could they take? |

*Note.* CL = Cognitive Load: expected demand based on Bloom's Taxonomy (Anderson et al., 2001), ranging from 1 = 'Low cognitive load demand' to 6 = 'High cognitive load demand'.

### Data pre-processing

After obtaining the answers from participants, the dataset was examined and cleaned. Some participants did not answer all six questions. Furthermore, among all received answers, 21 responses had less than 25 words or 140 characters. Previous research has shown that significantly small text samples can impact the performance of AV (Stein et al., 2007). However, there also have been effective AV practices with Twitter texts containing no more than 140 characters (Escalante et al., 2011), whose scheme for AV could be referred to. Therefore, as part of this study investigation, the dataset was tailored so that each text would need to have at least the length of a Twitter text. Twitter doubled the character limit from 140 characters to 280 characters in 2017, but in this study we followed the same approach presented in Escalante et al., (2011). As part of this process, we excluded all texts with less than 25 words (which is approximately 140 characters). Remaining texts were included in our analysis.

### Study 1: Validation of AV method with PAN14 dataset

After pre-processing our collected data, we developed the common character n-gram profile-based AV method proposed by Potha and Stamatatos (2014), which proved to be more effective under the circumstances where only short and limited numbers of sample texts are available. We then validated our implementation of the AV algorithm on a dataset retrieved from the PAN International Competition on Plagiarism Detection (Webis group, 2019a) so results could be compared with the ones published on Juola and Stamatatos (2013). PAN (Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection) is a series of scientific events and shared tasks on digital text forensics and stylometry (Meuschke and Gipp, 2013). They provide a series of openly shared text corpora for the scientific community to perform stylometric analysis and test AV methods for plagiarism detection. In this study, the "English Essays" test dataset from the 2014 PAN Competition (Webis group, 2019b) (referred to as "PAN14") will be used for validating our developed AV method. As shown in Table 2, PAN14 offered us a great dataset to validate our implementation as it provides several essays in

English. To perform this analysis, Study 1 was designed in a similar way to AV method presented in (Castro et al., 2015).

Moreover, previous studies based on common character n-gram profile-based AV method achieved fair results when tested on PAN14 corpus using 3-grams (Castro et al., 2015, Satyam et al., 2014). In this approach, n-grams are extracted without word boundaries, which means punctuation and blank spaces in the text are also included. They are good representation of writing styles of participants (Escalante et al., 2011). We followed the same approach as previous studies and used 3-grams in our investigation.

**Table 2: Statistics of the PAN14 authorship verification corpus**

| Corpus | Language | # Training documents | # Problems training | # Characters training (thousands) | # Test documents | # Problems test | # Characters test (thousands) |
|---|---|---|---|---|---|---|---|
| PAN14 | English | 729 (essays)<br>200 (novels) | 200 (essays)<br>100 (novels) | 3,450 (essays)<br>3,554 (novels) | 718 (essays)<br>400 (novels) | 200 (essays)<br>200 (novels) | 3,342 (essays)<br>13,772 (novels) |

*Similarity Functions*

Cosine similarity (referred to as "unknown similarity") between two count vectors (one from identified authors, another from an anonymous text) will be calculated and used as the classifier for verifying authorship, as shown in Figure 2. This approach is proposed by (Castro et al., 2015) and presented good results with character 3-gram features on PAN14 dataset. For comparison purposes, the metric for measuring the performance of our AV method is C@1 score (Penas and Rodrigo, 2011), which is also used in Stamatatos and colleagues (2014) for evaluating the participants' AV performances in PAN14. Once the performance of this AV method is evaluated and compared to other PAN14 participants using equivalent methods, we aim to apply the same method to our current collected data.

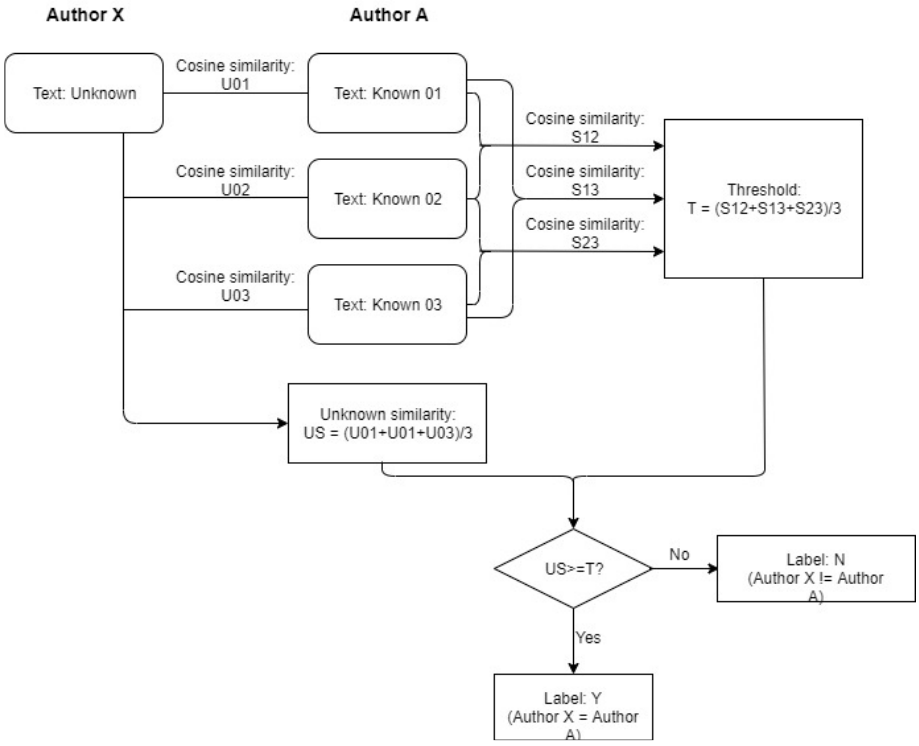


**Figure 2: Workflow of the authorship verification method applied to PAN14 dataset.**

## Study 2: AV with texts from same author

After validating the efficiency of our AV method with PAN14 dataset, we tested the efficiency of our AV method on our collected data. Only texts from the same participant were used in a single case for testing this AV method. This means that for a certain participant, two different pieces of texts by the same author were compared. In study 2, we did not compare texts from different participants. The structure of the current dataset in study 2 was designed in a slightly different way than the PAN14 dataset. We structured collected data in multiple folders. Each folder with an author ID contains six (or less) text files numbered 01-06 in accordance with the cognitive loads of their answered questions. To adapt to the current dataset and examine the impact of

cognitive load on the writings produced by a same author, this investigation was conducted as shown in Figure 3. As illustrated in Figure 3, to obtain a threshold for an author at a certain cognitive load (CL) level (n), this text was always compared against the text from the same author with CL 1, and a cosine similarity between these two were calculated and used as the threshold. Then, when compared with an anonymous text with a different cognitive load (m), the cosine similarity of these two (texts with CL m and n) were calculated and compared to the threshold to determine whether they were written by the same author.

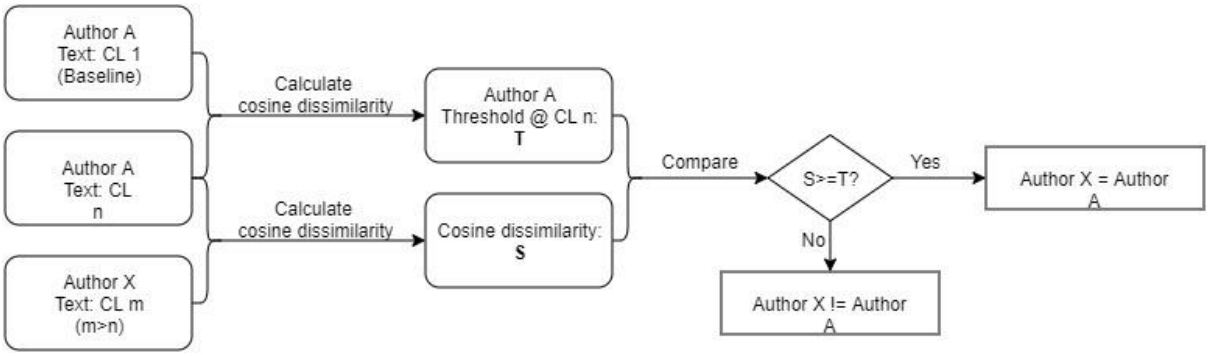


**Figure 3: Workflow of the authorship verification method applied to the current dataset; CL refers to the Cognitive Load of the question**

For each author, the process in Figure 3 was followed in each single AV case. To examine the impact of cognitive load on the participants' writings, comparisons were drawn between texts from different CL levels, as listed in Table 3. For example, in order to compare the texts of CL 2 and 3 from an author A, the threshold T was calculated as the cosine similarity between author A's answer to Q1 and author A's answer to Q2. Then, author A's answer to Q3 was regarded as an anonymous text and the cosine similarity S between this text and author A's answer to Q2 was then calculated. If the value of S was greater than or equal to T, this "anonymous text" was identified as written by author A; otherwise, it was regarded as written by a different author (i.e., fail to be correctly verified in this scenario). This process is referred to as "cross-CL level AV". For other cross-CL level AV listed in Table 3, the similar pattern was followed, with the author's answer to Q1 always used as a baseline for calculating the threshold T.

**Table 3: List of all categories of cross-CL level comparisons**

| Comparisons |
|---|
| CL 2 - 3 |
| CL 2 - 4 |
| CL 2 - 5 |
| CL 2 - 6 |
| CL 3 - 4 |
| CL 3 - 5 |
| CL 3 - 6 |
| CL 4 - 5 |
| CL 4 - 6 |
| CL 5 - 6 |

### Study 3: AV with texts from different authors

In Study 3, we tested the efficiency of our algorithm against texts produced by different authors. This is not a common practice for AV in academic context, but more of an exploratory attempt in our investigations. In this study, each author's writing was compared to all other authors' writings with a different cognitive load, with the AV process following the scheme of Figure 3 and comparisons categorised in the same cross-CL level AV process as before.

## Results and discussion

### Study 1: Validation of AV method with PAN14 dataset

Our AV method was first performed on the "English Essays" subset from the test dataset of PAN14 authorship

verification. The accuracy of our algorithm performance was calculated as C@1 = 0.580.
With reference to Stamatatos and colleagues (2014), the evaluated performances of the participants in the English Essays subset are presented in Table 4. Comparing our results with the ones from previous studies (Jankowska et al., 2013; Layton, 2014) who also employed common n-gram features and applied similarity distance as classifiers for AV of the same dataset, the C@1 score of our AV method was close to theirs (0.610 and 0.548), and also above the baseline score (0.530) presented for that dataset considering other submissions (i.e.: including other AV methods). The evaluation showed that this AV method achieved similar results as its equivalents and could be applied to our collected data.

**Table 4: Performance evaluation of PAN14 participants in English Essays**

| PAN14 Participants | C@1 Score |
|---|---|
| Layton, 2014 | 0.61 |
| Proposed AV method in this study | 0.58 |
| Jankowska et al., 2013 | 0.548 |
| BASELINE | 0.53 |

### Study 2: AV with texts from same author

After validating our AV method, we applied the AV algorithm to our collected data. In this part of the test, the process illustrated in Figure 3 was followed. The AV results were collected and the C@1 scores in each category of the comparison were calculated accordingly and presented in Table 5. AV performances in comparison CL 4-5 achieved the highest C@1 score of 0.941, while AV in CL 2-4 obtained the lowest C@1 score of 0.5. Considering the limited text sizes in the current dataset and the performance this AV method achieved in Study 1, it could be stated that regardless of the cognitive load changes in the texts, the AV method developed in this study could effectively identify writings from a same author. Furthermore, the results show that this AV method yielded higher C@1 score when at least one of the texts in the comparison correspond to a "Creative Work 2" (i.e. CL 5 or 6) question. This effect can be correlated with CL 5 and CL 6 responses having larger word counts average. 86% of answers for CL5 and CL6 questions in our study had between 100 and 300 words. However, the correlation between common character n-gram profile-based AV method accuracy and larger texts (over 500 words) might not be as straightforward and wasn't investigated in this study.

**Table 5: Evaluation results of cross-CL (Cognitive Load) level AV of texts from same author**

| Comparisons | C@1 score |
|---|---|
| CL 2 - 3 | 0.618 |
| CL 2 - 4 | 0.5 |
| CL 2 - 5 | 0.824 |
| CL 2 - 6 | 0.794 |
| CL 3 - 4 | 0.765 |
| CL 3 - 5 | 0.912 |
| CL 3 - 6 | 0.912 |
| CL 4 - 5 | 0.941 |
| CL 4 - 6 | 0.911 |
| CL 5 - 6 | 0.882 |

### Study 3: AV with texts from different authors

After examining the AV method on texts from a same author, we conducted comparisons between texts written by different authors. In this study, the AV process followed the scheme of Figure 3 and comparisons are presented in Table 6. Our findings show that C@1 scores obtained from these comparisons were lower than those from same-author comparisons, which means a great number of negative cases (i.e., two texts written by different authors) were incorrectly identified as positive (i.e., two texts written by the same author). This indicates the threshold set for the AV process was generally too low (i.e., lower than the similarity between two texts from different authors) to successfully identify a negative case.

To better understand obtained results and try to improve this performance, some statistical figures were obtained in terms of thresholds (T) and similarities (S) in this AV process. The difference (T − S) in each AV case was calculated and the mean value as well as standard deviation of them were derived from each category of the

comparisons, as listed in Table 6. It is noted that the standard deviation of T − S remained very stable around 0.1, regardless of the varied categories of cross-CL level comparisons.

We then experimented with increasing adopted threshold for determining authorship verifications in those scenarios. The original threshold obtained was increased by 0.104, which is a mean value of the standard deviations of all categories of the AV practice, as shown in Table 6. The verification processes remained the same. After making all the verifications, the C@1 scores were calculated again and listed in the rightmost column of Table 6. Compared to original C@1 scores, our new threshold significantly increased the accuracy of our comparisons.

**Table 6: Evaluation results of cross-CL (Cognitive Load) level AV of texts from different authors, before and after scaling up threshold values**

| Comparisons | C@1 Score | Difference = threshold ($T$) -cosine similarity ($S$) | | C@1 Score with scaled-up threshold ($T$+0.104) |
|---|---|---|---|---|
| | | Mean | Standard Deviation | |
| CL 2-3 | 0.448 | -0.01714 | 0.09357 | 0.798 |
| CL 2-4 | 0.534 | 0.01198 | 0.11070 | 0.853 |
| CL 2-5 | 0.288 | -0.0535 | 0.09383 | 0.721 |
| CL 2-6 | 0.396 | -0.02798 | 0.10111 | 0.764 |
| CL 3-4 | 0.314 | -0.04721 | 0.11529 | 0.673 |
| CL 3-5 | 0.117 | -0.10934 | 0.09519 | 0.466 |
| CL 3-6 | 0.177 | -0.09575 | 0.10744 | 0.511 |
| CL 4-5 | 0.135 | -0.09887 | 0.08742 | 0.511 |
| CL 4-6 | 0.182 | -0.08955 | 0.09866 | 0.546 |
| CL 5-6 | 0.272 | -0.08953 | 0.14134 | 0.54 |
| | | | Mean: 0.10446 | |

These results indicate that our AV method was not as accurate in identifying an author when comparing work of different authors. As an implication, the current paper supports use of stylometry for AV in higher education, particularly when comparing text written by the same student. This yields the need of creating leaner profiles database so individual learners' data can be stored and easily mined when required.

### Limitations and future improvements

Three limitations and possible directions for future work could be identified in this study. First, due to the limited text sizes, the AV methods that have proved to be effective in previous research, such as Unmasking (Koppel et al., 2007), could not be tested on the current dataset. Also, as there is only one piece of text available in each CL for each author, the cosine similarity calculated and adopted as threshold might be biased and not generalised enough for the verification process. If several texts in the same CL from one author could be collected, this threshold could be calculated as an average group similarity as illustrated in (Castro et al., 2015). Thus, it will be less biased and might achieve higher accuracy in the AV studies. Second, considering the limited number of participants in the data collection process, it remains an open question whether the AV method proposed in this study could be generalised and applied to a larger sample of academic writings. If data could be collected from a larger number of participants and tested with the current AV method, the results will be of stronger statistical significance. Lastly, cognitive load for each question was not measured, but rather, assumed based on previous research. To test this, future work could measure the actual cognitive load, for example through participants' self-reported cognitive effort.

## Conclusions

This study shows that authorship verification methods can provide good results to academic writings with varied cognitive loads. The results showed that with a valid AV method, the academic writings produced by students could be effectively verified. Findings also indicated that texts written by a same student could be successfully verified across different cognitive loads; moreover, when performing AV on texts of higher cognitive loads, the authorship is more likely to be successfully verified. This effect was found in responses with CL 5 and CL 6 as they had larger word counts average and richness of vocabulary. Larger responses supported better feature extraction and modelling students' (stylometric) profile.

These findings have important implications for the evaluation of academic integrity in higher education. Combined with anti-plagiarism tools such as Turnitin, AV methods can support educators identifying contract cheating. In this context, the use of AV in educational settings offer potential to enhance awareness around academic integrity issues beyond plagiarism, which can lead to better education around integrity issues. Moreover, in future, correlations between assessments' questions in different CL and frequency of AV issues in those can assist educators with assessment redesign.

## Acknowledgements

The authors would like to thank Xueying Lin for her contributions with some of the data analysis in the study.

## References

Abbasi, A., & Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, *26*(2), 1-29

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman,.

Awdry, R., Dawson, P., & Sutherland-Smith, W. (2022). Contract cheating: To legislate or not to legislate-is that the question?. *Assessment & Evaluation in Higher Education*, *47*(5), 712-726.

Beilock, S. L., & DeCaro, M. S. (2007). From poor performance to success under stress: working memory, strategy selection, and mathematical problem solving under pressure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(6), 983.

Brizan, D. G., Goodkind, A., Koch, P., Balagani, K., Phoha, V. V., & Rosenberg, A. (2015). Utilizing linguistically enhanced keystroke dynamics to predict typist cognition and demographics. *International Journal of Human-Computer Studies*, *82*, 57-68.

Calix, K., Connors, M., Levy, D., Manzar, H., MCabe, G., & Westcott, S. (2008). Stylometry for e-mail author identification and authentication. *Proceedings of CSIS research day, Pace University*, 1048-1054.

Castro, D. C., Arcia, Y. A., Brioso, M. P., & Muñoz, R. (2015, September). Authorship verification, average similarity analysis. In *Proceedings of the international conference recent advances in natural language processing* (pp. 84-90).

Escalante, H. J., Montes-y-Gómez, M., & Solorio, T. (2011, November). A weighted profile intersection measure for profile-based authorship attribution. In *Mexican International Conference on Artificial Intelligence* (pp. 232-243). Springer, Berlin, Heidelberg.

Holmes, D. I., & Kardos, J. (2003). Who was the author? An introduction to stylometry. *Chance*, *16*(2), 5-8.

Jankowska, M., Keselj, V., & Milios, E. (2013). Proximity based one-class classification with Common N-Gram dissimilarity for authorship verification task. In *CLEF 2013 Evaluation Labs and Workshop–Working Notes Papers* (pp. 23-26).

Juola, P., & Stamatatos, E. (2013). Overview of the Author Identification Task at PAN 2013. *CLEF (Working Notes)*, *1179*.

Kešelj, V., Peng, F., Cercone, N., & Thomas, C. (2003, August). N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING* (Vol. 3, pp. 255-264).

Koppel, M., & Schler, J. (2004, July). Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning* (p. 62)..

Koppel, M., Schler, J., & Bonchek-Dokow, E. (2007). Measuring Differentiability: Unmasking Pseudonymous Authors. *Journal of Machine Learning Research*, *8*(6).

Koppel, M., & Winter, Y. (2014). Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, *65*(1), 178-187..

Laramée, F. D. (2018). Introduction to stylometry with Python. *The Programming Historian*, *7*.

Layton, R. (2014). A Simple Local n-gram Ensemble for Authorship Verification. In *CLEF (Working Notes)* (pp. 1073-1078).

Macfarlane, B., Zhang, J., & Pun, A. (2014). Academic integrity: a review of the literature. *Studies in higher education*, *39*(2), 339-358.

Meuschke, N., & Gipp, B. (2013). State-of-the-art in detecting academic plagiarism. *International Journal for Educational Integrity*, *9*(1)..

Mosteller, F., & Wallace, D. L. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. *Journal of the American Statistical Association*, *58*(302), 275-309.

Oliveira, E., Conijn, R., De Barba, P., Trezise, K., van Zaanen, M., & Kennedy, G. (2020). Writing analytics across essay tasks with different cognitive load demands.Parkman, J. M., & Groen, G. J. (1971). Temporal aspects of simple addition and comparison. *Journal of Experimental Psychology*, *89*(2), 335.

Peñas, A., & Rodrigo, A. (2011). A simple measure to assess non-response. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (*p*p. 1415–1424), Portland, Oregon, USA. Association for Computational Linguistics.

Potha, N., & Stamatatos, E. (2014, May). A profile-based method for authorship verification. In *Hellenic Conference on Artificial Intelligence* (pp. 313-326). Springer, Cham.

Potthast, M., Braun, S., Buz, T., Duffhauss, F., Friedrich, F., Gülzow, J. M., ... & Hagen, M. (2016, March). Who wrote the web? Revisiting influential author identification research applicable to information retrieval. In *European Conference on Information Retrieval* (pp. 393-407). Springer, Cham.

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

Satyam, A., Dawn, A. K., & Saha, S. K. (2014). A statistical analysis approach to author identification using latent semantic analysis. *Notebook for PAN at CLEF*.

Stamatatos, E., Daelemans, W., Verhoeven, B., Potthast, M., Stein, B., Juola, P., ... & Barrón-Cedeño, A. (2014). Overview of the author identification task at PAN 2014. In *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014* (pp. 1-21).

Stein, B., Koppel, M., & Stamatatos, E. (2007, December). Plagiarism analysis, authorship identification, and near-duplicate detection PAN'07. In *ACM SIGIR Forum* (Vol. 41, No. 2, pp. 68-71). New York, NY, USA: ACM..

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive science*, *12*(2), 257-285.

Webis group (2019a). Pan shared tasks. https://pan.webis.de/shared-tasks.html. [Online; accessed May 2022].

Webis group (2019b). Pan14-verification. https://pan.webis.de/data.html. [Online; accessed June 2022].

Oliveira, E. & De Barba, P. (2022). The Impact of Cognitive Load on Students' Academic Writing: An Authorship Verification Investigation. In S. Wilson, N. Arthars, D. Wardak, P. Yeoman, E. Kalman, & D.Y.T. Liu (Eds.), *Reconnecting relationships through technology. Proceedings of the 39th International Conference on Innovation, Practice and Research in the Use of Educational Technologies in Tertiary Education, ASCILITE 2022 in Sydney*: e22177. https://doi.org/10.14742/apubs.2022.177