Forensic Assignment Stylometry

Robin Crockett.
Academic Integrity Lead, University of Northampton, UK.
Academic Visitor in Academic Integrity, Loughborough University, UK.
email: robin.crockett@northampton.ac.uk.

Abstract

This chapter discusses the stylometry of portfolios of assignments submitted by individual students from the perspective of evidence gathering in cases of suspected contract cheating. It opens with a discussion of the properties of student-submitted assignments that distinguish those from 'normal' texts as written by established authors, and how those affect stylometry. Stylometry has evolved with a main focus on established writers writing in their individual styles whereas students are learners developing their styles across a range of different types of assignments over their periods of study, often spanning several years. Any forensic stylometry of student assignments thus has to factor-in such variabilities while discriminating documents written by students from those written by commissioned third parties. The second part of the chapter describes a novel stylometric approach based on information theory, still in development, that seeks to work around variabilities in student assignments. Initial evaluative case-studies are presented.

Keywords: contract cheating, essay mill, ghost-writer, stylometry, information theory, cluster analysis, authorship attribution.

Introduction

The author has been researching stylometry of student-submitted assignments intermittently for the past decade, a period which has seen a close-to order of magnitude increase in the number of assignment providers, i.e., essay mills and assignment ghost-writers, with easy-to-find websites and social-media presences, accompanied by a close-to order of magnitude decrease in the cost of commissioning. Furthermore, over the past few years, the author and colleagues have observed an increasing proportion of submitted assignments that contain no forensically useful document metadata. The reasons for this are unclear but since the introduction of legislation to ban assignment providers in various countries (e.g., New Zealand in 2011, Ireland in 2019; Australia in 2020, England in 2022), it has become increasingly important that assignment providers take steps to minimise the chances that their student customers are caught and, thereby, minimise the chances that they themselves are identified and subjected to legal penalties. Also, there appears to be an ongoing shift away from students finding assignment providers' websites and transacting online, towards assignment providers proactively contacting students by social media, moving the transactions away from desktop/laptop computers to mobile devices accompanied by the use of web-based office software to provide commissioned assignments to students. Indeed, many students are contacted via their social media as soon as they leave secondary/high school, sometimes after they post "I'm heading to university..." social-media messages, and before they enrol at their higher education institutions. Furthermore, and sadly, some of those students will be experienced users of assignment providers by the time they enrol at their higher education institutions. Lastly, it should not be forgotten that it is, and always has been, easy for students with minimal engagement in their studies and/or minimising the chances of being detected to 'launder' commissioned assignments to remove at least headline incriminating document metadata.

Thus, it appears that academic misconduct investigators will have to become less reliant on document metadata and increasingly have to assemble bodies of evidence, including stylometric evidence, that form the bases of disciplinary cases. However, assignment stylometry is not straightforward owing to the nature of student assignments and the constraints imposed by assignment briefs as a necessary part of the learning and development of students.

To illustrate a core problem, consider the hypothetical example of a student who submits two essays and a technical report on related topics within the subject of study, and one of the essays is a commissioned ghost-written piece of work. On analysis, some stylometric approaches indicate that

the two essays are more similar to each other than either is to the report, whereas others indicate that the student-written essay and report are more similar to each other than either is to the ghost-written essay. Under circumstances where that investigation is triggered by a grading tutor's hunch (i.e., informed, intuitive opinion rather than hard evidence) that the three submissions cannot be the work of the same writer, if all the investigator has to work with are the three texts, with no external or document metadata information, how do they attribute authorship?

Under circumstances where there are other assignments that can be brought into the investigation and/or some external and metadata information, it can be possible to for example, use cluster-analysis techniques to identify those assignments which cluster with known student-written documents and/or those that cluster with known or suspected ghost-written documents. However, that takes time and can still leave the investigator with a not fully consistent set of statistical results to consider objectively in line with institutional, and legal, requirements as to what constitutes evidence of an acceptable nature and standard. For a concise consideration of such factors in a legal context, see Urbaniak and Bi Bello (2021).

It is issues such as these that prompted the author to look at possible different approaches and propose a novel approach, outlined herein, not as a single analysis that invariably produces a definitive attribution of authorship across a portfolio of student-submitted assignments but essentially as a triaging technique to inform further investigation.

Stylometry

Stylometry means, reasonably literally, 'measurement of style', i.e., statistical analysis of variation of (writing) style with authorship (e.g., Eder et al., 2016). Although stylometry can and does include analysis arising from direct (human) reading of texts, more generally in recent years it has come to mean analysis of (collections of) texts using computational techniques. Stylometry has developed in large part for the purpose of attribution of authorship of literary texts, or parts of literary texts. In that context, in light of the media coverage such research receives when it is published, many people, particularly in English-speaking countries, might think of the ongoing debates regarding the authorship of Shakespeare's works (e.g., Fox & Ehmoda, 2012; Aljumily, 2015; Wiggins, 2016). More notable perhaps as it relates to a living author, there is Juola's decisive identification of the pseudonymous author Robert Galbraith as being the Harry Potter author J K Rowling (Juola, 2013).

Stylometry generally involves statistical analysis of data directly derived from texts rather than texts themselves, such as word and *n*-gram frequencies (where *n* is general integer, e.g., a 2-gram is a two-word phrase, a 3-gram is a three-word phrase, and so on). Such approaches are sometimes referred to as 'bag of words' (e.g., Eder et al., 2016). There are also statistics on numbers of words and clauses per sentence, use of punctuation, and various metrics such as readability (Lines, 2022). Therefore, stylometry (increasingly computationally based) can reveal stylistic similarities and differences among texts that are not readily discernible to readers. However, stylometry is rarely a 'silver bullet' and, in general, it yields complex sets of statistical outputs which are not always fully consistent and which require careful, informed objective interpretation.

In the general case, stylometry of a collection of texts involves several iterations, each iteration involving various stages of data preparation and processing. Before any analysis can be undertaken, the collection of texts must be converted into a corpus (pl. corpora), i.e., a structured set of texts in a uniform (electronic) format. Once that initial stage has been completed and verified, it is probable that a corpus will be processed in various ways depending on the intended analyses. Such processing can typically include, for example, conversion of texts to sequences of individual words with some or all punctuation stripped out, possibly with conversion of all words to lower case. On occasion, more complex processing is undertaken, such as the removal of stop-words (also referred to as function-words, e.g., structural/grammatical words such as articles, conjunctions, prepositions, pronouns, particles) such that the text is reduced to content-words (also referred to as lexical-words, e.g., words such as main nouns and adjectives, main verbs and adverbs), or part-of-speech (POS) tagging (López-Escobedo et al., 2013).

Such techniques and processes have been established for consideration of portfolios of texts written by established, professional writers/authors, who are using language correctly as they intend, and writing with mature styles of their own choosing. This is not the case with students who write with varying styles depending on assignments that evolve over periods of study.

Stylometry of Student Assignments

Students write assignments according to assignment briefs: individual students can submit assignments in a range of styles over their periods of study including, for example, essays, reports, literature reviews, précis and dissertations. Assignments can be summative, i.e., graded to evaluate the learning and academic practice demonstrated by students (and contribute to the classification of the qualification being studied for), or formative, i.e., assessed by tutors in order to provide feedback to students to help them improve their learning and academic practice but with no grade. However, in the context of stylometry, assignments can be considered as ranging from concise reports, e.g., written with short 'staccato' sentences and pseudo-paragraphs of bullet-points with narrow vocabularies, to discursive essays, e.g., written with longer, multi-clause sentences in linked, fully developed paragraphs with broader vocabularies. Typically, students will submit assignments on a variety of topics, written in a variety of styles and formats, over their period of study. A further complication is that students are learners, and the expectation is that they develop their academic writing styles over the period of their studies in-step with their academic learning, meaning that, for example, a culmination-of-studies dissertation could have an observably more mature academic writing style than an early-stage submission.

In general, an assignment is on a set topic in a set format, often with detailed context and the approach to be taken further stipulated, sometimes leaving students little choice or flexibility. Thus, every student in the cohort is essentially trying to write the same perfect assignment in the same style and format to optimally address the task specified by the tutor. Furthermore, those students will have attended lectures, tutorials, seminars etc. in the topic area where they will have been exposed to the tutors' vocabularies, interpretations, etc. and (ideally) will have undertaken (some of) the recommended reading. That cohort might simultaneously have other different (types of) assignments on other topics set by other tutors, meaning that an individual student can produce and submit several markedly different assignments at set points in their programme of study.

Thus, a portfolio of assignments submitted by an individual student might include essays, reports and other formats including a dissertation (potentially less constrained by an assignment brief), across a range of topics within the programme/subject of study, all of which can influence or suppress aspects of individual writing style (Brocardo et al., 2013). This is the expectation regarding individual students even where there is no third-party authorship of assignments: any forensic stylometry of a portfolio of student-submitted assignments for possible third-party authorship of some of the submitted assignments must take account of such factors.

With regard to assignment providers, (some) ghost-writers take steps to make their products appear student-written in order to minimise the chances of detection during grading and assessment processes, others do not. Assignment ghost-writers range from, at one extreme, professional writers capable of writing very convincingly in student-like styles at a variety of levels, to, at the other extreme, essentially unqualified writers (sometimes themselves students) who lack such capability. Assignments are generally produced in rapid succession as efficiently as possible (to maximise output and income) and, perhaps, with subject-oblivious editing to minimise word-by-word similarity to deceive text-matching software (Crockett, 2022).

Furthermore, the author's and others' experiences acquired while investigating portfolios of student assignments indicates that it is not uncommon for commissioned assignments to earn poor grades or fail, meaning 'quality' is an unreliable discriminator. Also, it is sometimes the case that a student's own use of language is better and more consistent than that of the writer(s) they commission, even within individual assignments. In a related context, under some circumstances, e.g., dissertations and theses (with different writers taking distinct sections) and short-timescale commissions (where, in illustrative terms, two writers each write halves in parallel thereby halving

the time between commissioning and delivery compared to a single writer), several ghost-writers might work on single commissions leading to internally stylistically inconsistent documents.

At best, the complexity of forensic assignment stylometry is increased by such factors, at worst, such factors result in an ultimately inconclusive analysis that provides no useful authorship information.

Approaches to Forensic Assignment Stylometry

Having considered the complicating factors, and before considering what is reasonably possible, let us consider what the reasonable objectives of a forensic stylometric investigation of a portfolio of assignments submitted by a student are, or should be. In essence, the objective is not to identify one or more specific ghost-writers – sometimes ghost-writers are identifiable, but this should be considered as a bonus rather than the norm – but is to identify two or more subsets/clusters of assignments according to writing style that when considered as a whole, on balance of probabilities, do not align with having been written by a single individual and so align with multiple authorship. Clearly, care must be taken with group-work and other assignments that necessarily contain material written by others. Where this objective is met, it means that while the student in question might have written the assignments in one subset/cluster, on balance of probabilities they can't also have written the others in the other subset(s)/cluster(s).

Following the stylometry, it is probable that human inspection/reading will be needed to identify specific tell-tales, and some of these might be obvious even without forensic stylometry, and can include, for example (Crockett, 2022), consistent differences in:

- words that have different spellings, e.g., in English, 'gaol' vs. 'jail', 'judgment' vs. 'judgement', 'adviser' vs. 'advisor',
    - as well as American vs. British English spelling conventions,
    - and equivalents in other languages;
- capitalisation of proper nouns;
- punctuation and grammar, including contractions and eclipsis;
- phrasing and word use;
- errors in spelling, punctuation and grammar
    - with the proviso that many/most students and ghost-writers use built-in spelling-checkers and increasingly use grammar-checkers.

With regard to that last point, individual misuses of grammar (including punctuation, capitalisation of proper nouns etc.) sometimes help to identify potential differences in authorship across a student's portfolio of submitted assignments, as previously reported (Crockett & Best, 2020). However, this is a decreasingly useful approach due to the increasing use of spelling and grammar checkers and online tools which have such functionality built-in (Rogerson & McCarthy, 2017). Of greater concern, perhaps, is the growing availability of artificial intelligence (AI) assisted tools such as GPT-3 (Johnson & Isiev, 2022) and, latterly, ChatGPT, GPT-4 and Bard, including 'essay bots' (which, in the current context, can be considered as AI ghost-writers). The use of such tools, which are also used by assignment providers, complicates what have previously been reliable differentiators of authorship.

As well as reader-identifiable tell-tales that can help complete the evidence obtained from forensic stylometry, in some circumstances there will be other information, e.g., a known/identifiable writer-name in the metadata or a document provided by a whistle-blower, that identifies one or more assignments as ghost-written. Under such circumstances, stylometry can reveal subsets/clusters of assignments that accord with that information and, possibly, identify other assignments that align with styles other than the student's. Of course, under other circumstances, there will be no such metadata or external information and stylometry will be inconclusive, revealing no consistent subsets/clusters among the assignments.

For an overview of stylometric approaches to the identification of contract-cheated work see Juola (2017). However, here are some basic pointers for forensic stylometry of portfolios of student-submitted assignments, based on the author's research, that should help minimise false positives (and false negatives) for third-party authorship (with specific regard to English):

- analyse both case-preserved and all-lower-case versions of the assignments;
- preserve hyphenation and apostrophes;
- analyse punctuation as well as word and *n*-gram frequencies;
- retain stop-words;
- possibly exclude content-words that are highly specific to individual assignment topics.

On occasion, an investigator is lucky: they employ all reasonable/justifiable stylometric analyses on a portfolio of student-submitted assignments and all yield the same subsets/clusters such that the overall investigation gives a clear and unambiguous indication regarding authorship of all the assignments. In the author's experience, that very rarely happens. More commonly, the employed analyses will yield the same core subsets/clusters but with some assignments showing varying subset/cluster membership such that the investigation gives a clear indication regarding the authorship of some but not all assignments. However, in some circumstances, the analyses might yield mutually exclusive results, with no consistent core subsets/clusters, resulting in an inconclusive investigation.

Ideally, what is needed is a stylometric technique that is sensitive to the significance of words in texts, ratios of high- and low- significance words, patterns in the usage of words, even where words of similar significance vary among the texts in a corpus, e.g., according to assignment topic. A technique with such properties would have the potential to 'see' through some of the variations in style within a portfolio of student-submitted assignments where a student has necessarily varied their writing style and vocabulary to address assignment briefs. Whether such a technique is realisable is an open question, but there are some approaches that can partly address some of the issues, and the following section presents a snapshot of the author's current research.

Information Entropy and Self-information

The approach reported here stemmed from the author's previous time-series analysis (TSA) research and started by considering a text as an ordered sequence (cf. time-series) of words (or *n*-grams), in effect, considering a text as a categorical time-series. The initial objective had been to investigate texts for repeated and periodic features. However, the requirements of having to undertake contract-cheating investigations of increasing scale and complexity have necessitated the more immediate objective of developing a rapid stylometric triaging tool that provides a robust preview of any clustering/sub-setting of assignments.

A basic underpinning requirement is how to numerically encode the words. Noting that a text can be considered as a sequence of relatively few (low probability) high-significance content-words which carry the core meaning, interspersed with relatively many (high probability) low-significance stop-words which expand and qualify the core meaning, an encoding that reflects these basic properties was sought. The underpinning assumption is that individual authors use consistent patterns of high- and low- significance words across different topics and document types. Thus, an encoding of this type has the potential to identify similarities and differences in style at sub-sentence-level, even where the actual words differ.

In a previous paper (Crockett & Best, 2020), the author reported a prototype approach which involved encoding words as their numbers of syllables on the basis that, in English, high-probability (stop-) words tend to be single-syllable whereas low-probability (content-) words tend to be multi-syllable. Despite the simplicity of the approach, because stop-words and content-words do not always conveniently align with numbers of syllables, this approach proved to be sufficiently informative to progress to the more rigorous information-entropy approach reported herein.

This approach involves representing words (or *n*-grams) by their self-information (SI, also referred to as surprisal or Shannon information), and analysing the transformed texts. SI can be considered as a measure of textual significance of a word determined by that word's probability of occurrence in the text or corpus. In his pioneering analysis, Shannon (1948) set out the framework for information entropy in the context of data-communication systems and transmission bandwidths. Shannon defined the entropy of (a body of) text as the expectation of the self-information of the symbols used to represent that text, with the symbols in this context being binary representations of letters and punctuation marks.

Therefore, slightly adapting Shannon's notation, the self-information, *SI*, of a symbol *x* in a (body of) text, *X*, is defined as

$$SI(x)=\log_2(\frac{1}{p(x)})=-\log_2(p(x))$$

where *p(x)* is probability of symbol *x* occurring in *X* (and noting that Shannon used base-2 logarithms because he was considering binary-encoded systems).

From there, the entropy, *H*, of the overall (body of) text *X* is defined as

$$H(X)=\sum_{x\in X} p(x)SI(x)$$

i.e., the entropy is the expectation of the self-information.

Shannon's analysis can be adapted for TSA and stylometry purposes, with the essential modification being that the symbols are words rather than letters, and the overall body of text is a corpus of texts, e.g., as prepared from a portfolio of student-submitted assignments. Also, as we are not considering binary-encoded systems in this context, there is neither necessity nor advantage in using base-2 logarithms and, provided an analysis is fully consistent, natural or base-10 logarithms are much more convenient. Indeed, for stylometric purposes, it is not necessary to take logarithms but the term self-information (SI) is retained because it describes the core concept in this context, and the same results are obtained whether or not logarithms are taken (see following section).

This is comparable to the TF-IDF (term frequency – inverse document frequency) approach used in natural language programming (NLP) where IDF is a measure of the proportion of documents in a corpus that contain a given word and TF is the total frequency of that word in a document (Robertson, 2004).

Stylometry via Self-Information

The most accessible ways of analysing the transformed text are to perform cluster analysis (as described herein) or principal components analysis (PCA) on SI frequencies, analogous to the existing stylometric technique for analysing word (*n*-gram) frequencies. Therefore, a convenient starting point is a table of word (*n*-gram) frequencies, i.e., the starting point for word (*n*-gram) frequency analysis, case-preserved or lower-case and accounting for or ignoring 'in-word' punctuation such as hyphens, apostrophes etc. as appropriate. The underpinning assumption for this is that it is the SIs that are important, just as it is the words that are important in conventional word-frequency analysis, and not the grammar and punctuation (Eder et al., 2016).

That table of word (*n*-gram) frequencies is transformed to a table of SI frequencies: the total word (*n*-gram) frequencies across all texts are converted to SIs (via probabilities, noting that SIs are uniquely defined by probabilities) and, where words have the same SI (probability), those columns (rows) are combined such that each SI (probability) has one column (row) of frequencies across all texts in the corpus. Thus, in general, a table of SI frequencies will have fewer columns (rows) than the table of word frequencies it is derived from and, at most, will have the same number.

In the following examples, the analysis is (agglomerative) hierarchical cluster analysis of the SI frequencies, and all SI frequencies are used. All the analysis was performed using the R open-source statistical computing software (R Core Team, 2022) with the add-in packages Stylo (Eder et al., 2016) and Pvclust (Suzuki & Shimodaira, 2006). The clustering metrics used are Euclidean distance and the Ward-D2 clustering algorithm (Ward, 1963; Murtagh & Legendre, 2014), and Pvclust calculates p-values using an approximately unbiased (AU) multi-scale bootstrap resampling algorithm, with 10,000 bootstrap iterations in the following examples.

Expressed informally, the p-value is the probability that the null hypothesis applies, assuming the null hypothesis is valid. The null hypothesis here is that the SI frequency lists at each level in the clustering are samples from a common source, with the over-arching hypothesis that SI frequencies associate with writing style and authorship. The clusters are determined by the similarities among the SI frequency lists according to the metric used, the p-values are calculated following cluster determination.

<u>Illustrative example: nine eighteenth-century novels</u>

The R Stylo software includes a dataset/corpus of nine novels as follows (labels used in the results in brackets): *Emma* (Em_JA), *Pride and Prejudice* (PP_JA), *Sense and Sensibility* (SS_JA) by Jane Austen; *Jane Eyre* (JE_CB), *The Professor* (Pr_CB), *Villette* (Vi_CB) by Charlotte Brontë; *Agnes Grey* (AG_AB), *The Tenant of Wildfell Hall* (TW_AB) by Anne Brontë; and *Wuthering Heights* (WH_EB) by Emily Brontë. It is emphasised that this is not a literary analysis, and this corpus is chosen solely because it is fixed and available to anyone who wishes to reproduce the analysis without them having to obtain their own, possibly different, versions of the texts from different (online) resources. The Stylo authors state these texts are, quoting directly from the software, 'harvested from open-access resources, e.g. the Gutenberg Project' (Gutenberg, n.d.).

Results: Explanation of Dendrograms

Figure 1 shows the output of the analysis. This is a vertical dendrogram, i.e., a representation of the relationships revealed by a hierarchical cluster analysis (Eder, 2012). With reference to Figure 1 (and the other figures), the dendrogram shows a set of horizontal 'edges' or 'nodes' (i.e., clusters, cluster levels), each containing vertical 'branches' or 'clades' (i.e., cluster members and sub-clusters) all comprised of individual 'leaves' (i.e., corpus members, labelled at the foot of the figure). Each edge is labelled with the AU p-value (in percent) for the cluster membership at that level. Within a cluster, the greater the vertical distance between the upper and lower edges, the more distinct the branching at the upper level. Thus, the further down that a cluster diverges, the greater the similarity of the members within it relative to similarity to other clusters.

< Figure 1 here >

Figure 1. Clustering of the Nine Eighteenth-Century Novels.

With reference to Figure 1 (and the other figures), the clustering algorithm iterates through the corpus SI frequency lists, initially considering each corpus member as a single-element cluster, and reducing the number of clusters by one at each iteration as the most similar corpus members are successively identified. On the first iteration, the algorithm identifies the two most similar members, i.e., *The Professor* and *Villette*, corresponding to the lowest edge/node. It then treats those two as a two-member cluster and iterates through the eight remaining clusters to identify the next most similar pair, i.e., *Agnes Grey* and *The Tenant of Wildfell Hall*, corresponding to the next lowest edge/node. It combines those as a two-member cluster and repeats the process iteratively, corresponding to the successively higher edges/nodes, until it terminates at one whole-corpus cluster. That is always the final stage in agglomerative hierarchical cluster analysis and does not convey any clustering information. Thus, in this example:

- the third iteration identifies the cluster *Pride and Prejudice* and *Sense and Sensibility*;
- the fourth iteration clusters *Jane Eyre* with the other two Charlotte Brontë novels;
- the fifth iteration clusters *Emma* with the other two Jane Austen novels;
- the sixth iteration clusters *Wuthering Heights* (Emily Brontë) with the two Anne Brontë novels;
- the seventh iteration clusters all six Brontë Sisters novels;
- the eighth (final) iteration combines the Jane Austen and Brontë Sisters novels into the final whole-corpus cluster.

Results: Interpretation

Working down from the whole-corpus cluster, the analysis indicates that the three Jane Austen novels are distinct from the six Brontë Sisters novels, each cluster with p=100%. Moving down within the Brontë Sisters cluster, the next level clusters Charlotte's novels with p=82%, distinct from Emily's and Anne's novels which cluster with p=85%. Within that (sub-)cluster, Anne's novels cluster with p=98%, distinguishing those two from Emily's novel. These clusters are all strong null hypothesis, indicating strong clustering according to SI frequencies and, therefore, implying strong clustering according to authorship.

Before we consider two student-submitted portfolios, let us consider what we might deduce, if, for example, we had the texts of *Emma* or A*gnes Grey* but our only authorship information was that these were written by either Jane Austen or one of the Brontë Sisters. In such circumstances, this analysis would clearly indicate Jane Austen as most probable author of *Emma* and one of the

Brontë Sisters as most probable author of *Agnes Grey*, with Anne or Emily more probable than Charlotte. It is emphasised that in both these hypothetical cases, or in cases where we had other texts to consider, this analysis of itself is not a definitive determination of authorship and further analysis combined with other information would be needed to increase the confidence of authorship attribution.

In both the following examples, the texts were prepared by removing the reference lists and all student name/identity information from the submitted assignments.

Example 1

This example comprises a portfolio as described in Crockett & Best (2020) and, briefly, the academic integrity investigation was initiated in response to information from the police that the student in question had made a succession of payments to an identified assignment provider. The portfolio comprises 20 submitted assignments labelled as follows:

- prefix either L6 (final-year bachelors) or L7 (masters);
- two-letter single-number code indicating topic and submission sequence within that topic;
- those assignments with evidence of ghost-writing, i.e., third-party authorship, suffixed according to sources as described below.

The five assignments suffixed 'A' were commissioned from the provider identified by the police. That includes the two assignments suffixed 'AB' previously attributed to identified ghost-writer 'B' (Crockett & Best, 2020), who has subsequently been identified (via unrelated online legal reporting) as being retained by that provider, and probably two writers using the same writer-name. The assignments suffixed 'C' and 'D' were written by identified ghost-writers possibly independent of that provider. The masters dissertation is suffixed 'X' and was identified as ghost-written by the supervisor and moderator ('DI' indicates dissertation and intermediate assignments, and note that the L6 and L7 dissertations are on different topics). The two submissions suffixed 'Y' and 'Z' contain metadata that do not explicitly identify assignment providers but which are otherwise consistent with third-party authorship.

Results

< Figure 2 here >

Figure 2. Clustering of Student Assignments, Example 1.

Figure 2 shows the output of the analysis. This clearly shows two top-level clusters (p=96%). One cluster of four assignments containing the 'AB2', 'C', 'Y' and 'Z' assignments, supporting the metadata evidence regarding the 'Y' and 'Z' assignments and indicating that all four are probably associated with the identified assignment provider. The other cluster of 16 assignments, including:

- a cluster (p=76%) of the five dissertation (DI) assignments, two L6 and three L7, indicating the possibility that both dissertations were written by the same assignment provider 'X';
- a cluster (p=76%) comprising
  - a sub-cluster (p=93%) of five assignments consisting of a sub-cluster of four assignments commissioned from provider 'A' plus one other assignment, L6_SP1, indicating a strong probability that L6_SP1 was also commissioned;
  - a sub-cluster (p=98%) of six assignments comprising two sub-clusters of three assignments, one of which contains the assignment from ghost-writer 'D'.

Prior to the stylometry, the five other assignments in that (sub-)cluster with the 'D' assignment had been considered as probably student-written (e.g., ambiguous metadata, general quality of content and presentation), but the clustering with the 'D' assignment raises questions regarding the authorship of those assignments, particularly L7_NS1 and L7_CS1.

Summary

This triage analysis has revealed similarities in the SI frequency lists, i.e., frequencies of high and low SI words that:

- confirm the initial information provided by the police,
- confirm known relationships between identified assignment providers and suggest others;
- indicate probable common third-party authorship of both dissertations and associated intermediate assessments;

- indicate possible third-party authorship of other assignments.

This analysis identifies the same core clusters previously indicated by extensive word-frequency analyses as reported in Crockett & Best (2020), but with some differences. This serves to illustrate the complexities inherent in stylometry where there are different combinations of stylistic similarities (and differences) among the corpus members, revealed by different analyses. In essence, comparing with the previous detailed word-frequency analyses, the common core-clusters around identified ghost-writing are (with reference to the clustering reported above):

- the five-assignment cluster L6_ET1.A1, L6_LL1.A2, L6_EL1.A3, L6_EL2.AB1, L6_SP1;
- the four dissertation assignments L6_DI2, L6_DI1, L7_DI2, L7_DI3.X,
  - the L7_DI1 assignment clusters inconsistently in the word frequency analyses;
- the three assignments L7_OC1.AB2, L6_SP2.C, L7_OC2.Z,
  - the L6_HR3.Y assignment clusters inconsistently in the word frequency analyses;
- the five assignments associated with the 'D' assignment cluster inconsistently in the word frequency analyses.

In summary, with regard to differences in authorship, we can have high degrees of confidence in the core-clusters that are revealed consistently across the individual analyses, and lower degrees of confidence in the other clusterings that are revealed less consistently.

Example 2

This example comprises a portfolio of ten assignments submitted by an individual student in a single academic year. The academic integrity investigation was initiated by tutors and moderators reporting clear variations in quality, presentation and style among the submissions that are strongly indicative of multiple authorship. That investigation revealed six assignments containing clear document properties information indicating three assignment providers, i.e., three assignments suffixed 'A', two suffixed 'B' and one suffixed 'C' in the results (different 'A', 'B' and 'C' to the previous example). The two-letter single-number code indicates topic and submission sequence within that topic (different labelling to the previous example, although 'DI' again denotes dissertation).

Results

< Figure 3 here >

Figure 3. Clustering of Student Assignments, Example 2.

Figure 3 shows the output of the analysis. This clearly shows two top-level clusters, one (p=84%) of eight assignments including the six assignments identified as ghost-written, and one (p=86%) of two early-stage assignments, possibly written by the student. Within the eight-member cluster:

- a sub-cluster (p=91%) of five assignments, two each by assignment providers 'A' and 'B' and another assignment, ML_1, indicating
  - providers 'A' and 'B' are possibly ghost-writers working for the same assignment provider and writing in a house style,
  - the strong probability that ML_1 was also written by 'A' or 'B';
- a sub-cluster (p=95%) of three assignments, one each by assignment providers 'A' and 'C' and another assignment, EL_1, indicating
  - providers 'A' and 'C' are possibly ghost-writers working for the same assignment provider and writing in a house style,
  - the strong probability that EL_1 was also written by 'A' or 'C'.

Summary

This analysis supports the evidence revealed by the academic integrity investigation and, further, suggests possible associations between writers 'A', 'B' and 'C' via the sub-clusters within the eight-member cluster. Also, the two other assignments in this eight-assignment cluster, i.e., ML_1 and EL_1, are implicated. This would require further investigation in order to provide evidence for an academic misconduct hearing. However, preliminary word-frequency analysis (not reported here) confirms the core top-level two-member/eight-member clusters, the clustering of the

six ‘A’, ‘B’ & ‘C’ assignments and the association of the ML_1 and EL_1 assignments with those six.

In essence, the SI analysis, supported by preliminary word-frequency analysis, indicates that eight assignments are probably ghost-written and two are possibly student-written, with six of the eight having clear document properties and metadata evidence highly consistent with assignment providers.

Future Work

There is much that remains to be done with regard to both how to undertake forensic stylometry of student-submitted assignments and the robust analytical techniques required for such investigations.

With regard to the former, the author and colleagues have observed that it is becoming more common for several students on the same course to use the same assignment provider. This might appear to be initially surprising, and an obvious conflict of the individual students’ interests they should avoid. However, it is attributed to the increase in assignment providers using social media and penetrating student social-media groups, often initially posing as students to gain access. This means it is likely that students in a social-media group use the same assignment provider independently and unaware of the others simply because they’ve been contacted by that provider and don’t discuss their dishonest behaviour.

A consequence of this change is that, whereas the focus herein has been portfolios of assignments submitted over time by individual students, it is necessary to develop equivalent guidelines for the investigation of portfolios of assignments submitted by course-cohorts both at the same point in time and over successive submissions of the same assignment. It is anticipated that guidelines for stylometry of course-cohort submissions will be more straightforward because, in such circumstances, all the submissions are (or should be) of the same type in response to the same assignment brief, with students at the same point in their academic development, i.e., corpora lacking the multi-style and time-line aspects of the circumstances considered herein.

The SI-based approach to stylometry has produced promising initial results, including very preliminary differentiation of AI-generated and human-written texts not reported here, and development and evaluation are ongoing. In particular, noting that over the period of preparing this chapter, the ChatGPT, GPT-4 and Bard AI tools have been released, development of the potential for SI-based, and other, stylometric approaches to distinguish AI-generated texts from human-written ones is of increasing importance.

On a related note, there are unexplored TSA approaches which might, for example, yield ‘waveform signatures’ that associate with individual writing styles, and thereby offer a further stylometric approach. There is some published work on recurrent features in texts (e.g., Coco & Dale, 2014; Allen et al., 2017), but that work has been based on words rather than types of words as represented by SIs.

SI representation of words is not the only possible approach to the problem. POS-tagging approaches offer the potential to do similar analysis (Szwed, 2017). However, POS tags do not reflect the relative frequencies whereas SI, or SI-related, representations of texts do, indicating that SI representation is possibly more sensitive to patterns of word use.

Summary

Factors to consider when undertaking forensic stylometry of portfolios of assignments submitted over time by individual students have been discussed. Some of these could be regarded as obvious, others less so, but all are important. The over-arching advice, irrespective of specific circumstances, is for an investigator to carefully consider what would be expected in light of the types of assignment submitted assuming student authorship throughout, and to account for that expectation when interpreting the stylometry results. Even where that is done and, for example, an investigator knows to expect certain similarities due to similarities among assignment briefs, it is probable that stylometry will yield some inconsistent results. In essence, the aim of forensic stylometry is to provide evidence, not proof, and that evidence needs to be considered objectively:

any unjustifiable assumptions or criteria increase the risk of obtaining false positives/negatives with obvious consequences for any disciplinary action which uses the flawed stylometric evidence.

The author and colleagues have observed an increase in the proportion of assignments that contain little or no forensically useful metadata. The reasons for this are unclear but, along with the increasing use of social media by assignment providers to transact commissions, there is as increasing use of web-based office software by assignment providers to provide commissioned work to students, and files downloaded from such software can contain little or no metadata. Also, students themselves use such software, meaning that honestly-written assignments can show similar presences/absences of metadata, albeit with entirely different motivations. In addition, most desktop software allows users to make privacy settings that essentially redact author metadata. Lastly, as higher education institutions continue to become increasingly aware of the forensic potential of document metadata, assignment providers will increasingly take steps to reduce the risks to their customers and themselves of such information being present and forensically useful.

The overall consequence of such trends and changes is that it is probable that contract-cheating investigations will have to increasingly include forensic stylometry as a means of obtaining evidence. This, in turn, necessitates research and development of analytic techniques.

References

Allen, L., K., Likens, A. D., McNamara, D. S. (2017) Recurrence Quantification Analysis: A Technique for the Dynamical Analysis of Student Writing. *Proc. Thirtieth International Florida Artificial Intelligence Research Society Conference*. Association for the Advancement of Artificial Intelligence (www.aaai.org).

Aljumily, R. (2015), Hierarchical and Non-Hierarchical Linear and Non-Linear Clustering Methods to 'Shakespeare Authorship Question'. *The Social Sciences*, 4, 758-799.

Brocardo, M., Traore, I., Saad, S. & Woungang, I. (2013). Authorship Verification for Short Messages Using Stylometry, *Proc. IEEE Intl. Conference on Computer, Information and Telecommunication Systems* (CITS 2013), 1-6. Athens, Greece, 7-8 May 2013.

Coco, M., I. & Dale, R. (2014) Cross-recurrence quantification analysis of categorical and continuous time series: an R package. *Frontiers in Psychology*. 5, 510, p1-14. doi:10.3389/fpsyg.2014.00510.

Crockett, R., G., M. & Best, K. (2020) Stylometric Comparison Of Professionally Ghost-Written And Student-Written Assignments. *Integrity in Education for Future Happiness*. 35-49. Mendel University Press. ISBN: 978-80-7509-772-9. doi:10.11118/978-80-7509-772-9-0035

Crockett, R., G., M. (2022) Presentation, Properties and Provenance: the three Ps of identifying evidence of contract-cheating in student assignments. In *Contract Cheating in Higher Education: Global Perspectives on Theory, Practice, and Policy* (Eaton, S., E.; Curtis, G., J.; Stoesz, B., M.; Clare, J.; Rundle, K. & Seeland, J. Eds.) Springer. ISBN: 9783031126796

Eder, M. (2012) Computational stylistics and biblical translation: how reliable can a dendrogram be? In *The Translator and the Computer* (Piotrowski, T., Grabowski, L. eds). 155–170. WSF Press, Wroclaw, Poland.

Eder, M., Rybicki, J., Kestemont, M. (2016) Stylometry with R: A Package for Computational Text Analysis. *The R Journal*, 8:1, pages 107-121. doi:10.32614/RJ-2016-007, ISSN: 2073-4859.

Fox, N. P. & Ehmoda, O. (2012). Statistical Stylometrics and the Marlowe-Shakespeare Authorship Debate. cs.brown.edu/research/pubs/theses/masters/2012/ehmoda.pdf (accessed 14/08/2022)

Gutenberg (n.d.) Project Gutenberg, www.gutenberg.org.

Johnson, S. & Iziev, N. (2022). A.I. Is Mastering Language. Should We Trust What It Says?. *The New York Times*. (April 15, 2022) www.nytimes.com/2022/04/15/magazine/ai-language.html

Juola, P. (2013) How a Computer Program Helped Show J. K. Rowling wrote A Cuckoo's Calling. *Scientific American*. 20 August 2013, Springer Nature.

Juola, P. (2017) Detecting Contract Cheating via Stylometric Methods. *Proc. Plagiarism across Europe and Beyond*. 187-198. Brno, Czech Republic. 24-26 May 2017.

Lines, N. A. (2022) The Past, Problems, and Potential of Readability Analysis. *CHANCE*, 35:2, 16-24, doi:10.1080/09332480.2022.2066411.

López-Escobedo, F., Méndez-Cruz, C-F., Sierra, G., Solórzano-Soto, J. (2013) Analysis of Stylometric Variables in Long and Short Texts. P*rocedia - Social and Behavioral Sciences*. 95, p604 – 611. doi: 10.1016/j.sbspro.2013.10.688

Murtagh, F. & Legendre, P. (2014) Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?. *Journal of Classification*, Springer. 31, 274–295. doi: 10.1007/s00357-014-9161-z

R Core Team (2022). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. www.R-project.org.

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*. 60 (5): 503–520. doi:10.1108/00220410410560582.

Rogerson, A. M. & McCarthy, G. (2017). Using Internet based paraphrasing tools: Original work, patchwriting or facilitated plagiarism? *International Journal for Educational Integrity*. 1833-2595. doi:10.1007/s40979-016-0013-y

Shannon, C. (1948) A Mathematical Theory of Communication. *Bell System Technical Journal*. 27 (3): 379-423.

Suzuki, R. and Shimodaira, H. (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22 (12): 1540-1542.

Szwed, P. (2017) Authorship Attribution for Polish Texts Based on Part of Speech Tagging. In: *Beyond Databases, Architectures and Structures. Towards Efficient Solutions for Data Analysis and Knowledge Representation* (Kozielski, S., Mrozek, D., Kasprowski, P., Małysiak-Mrozek, B., Kostrzewa, D. eds). BDAS 2017. Communications in Computer and Information Science, vol 716. Springer. doi:10.1007/978-3-319-58274-0_26

Urbaniak, R. and Di Bello, M. (2021) Legal Probabilism, *The Stanford Encyclopedia of Philosophy*, Ed. Zalta, E. N. Metaphysics Research Lab, Stanford University. Fall 2021. https://plato.stanford.edu/archives/fall2021/entries/legal-probabilism

Ward, J. H. Jr. (1963) Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, Taylor & Francis. 58:301, 236-244, doi: 10.1080/01621459.1963.10500845

Wiggins, M. (2016) Who Wrote Shakespeare? BBC. www.bbc.co.uk/programmes/articles/211LBPTmBYp2rbh4bSQlSTS/who-wrote-shakespeare

Fig 1

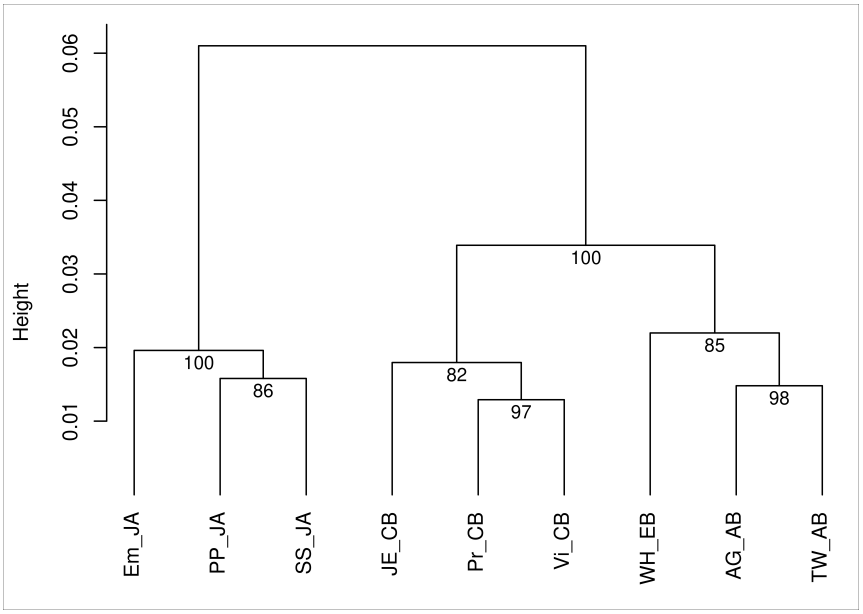

Fig 2

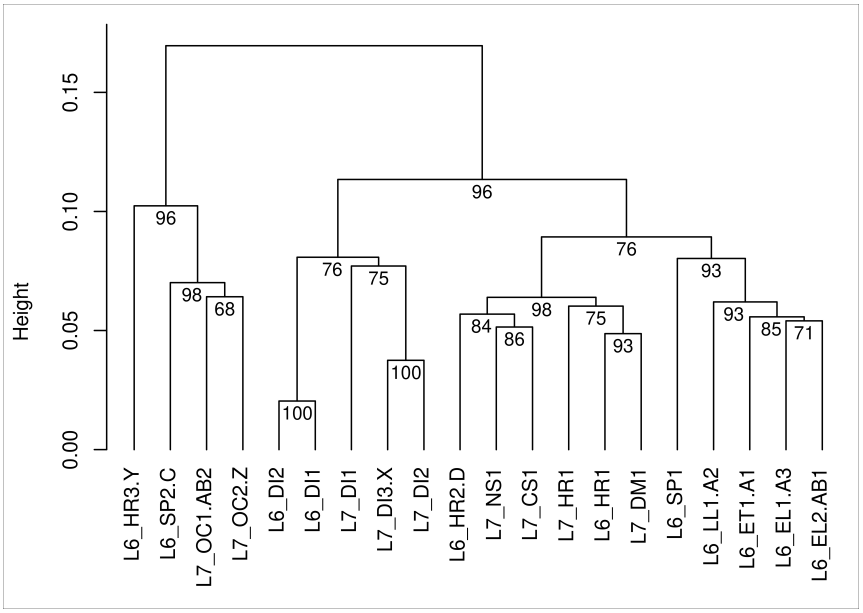

Fig 3