Evaluating Binary n-gram Analysis For Authorship Attribution

Mark Carman *mark.carman@mymail.unisa.edu.au*
School of Information Technology and Mathematical Sciences
University of South Australia
Adelaide, SA 5095, Australia

Helen Ashman *helen.ashman@unisa.edu.au*
School of Information Technology and Mathematical Sciences
University of South Australia
Adelaide, SA 5095, Australia

Abstract

Authorship attribution techniques focus on characters and words. However the inclusion of words with meaning may complicate authorship attribution. Using only function words provides good authorship attribution with semantic or character n-gram analyses but it is not yet known whether it improves binary n-gram analyses.

The literature mostly reports on authorship attribution at word or character level. Binary n-grams interpret text as binary. Previous work with binary n-grams assessed authorship attribution of full texts only. This paper evaluates binary n-gram authorship attribution over text stripped of content words as well as over a range of cross-domain scenarios.

This paper reports a sequence of experiments. First the binary n-gram analysis method is directly compared with character n-grams for authorship attribution. Then it is evaluated over three forms of input text, full text, stop words and function words only, and content words only. Subsequently, it was tested over cross-domain and cross-genre texts, as well as multiple-author texts.

Keywords: Authorship Attribution, Binary n-gram, Stop Word, Cross-domain, Cross-genre.

1. INTRODUCTION

Authorship attribution is the forensic analysis of text with the aim of discovering who wrote it. There is a range of authorship attribution methods in use, along with some refinements to their use intended to improve their accuracy. This paper describes a recent method, binary n-gram analysis, that is a variation of the well-established character n-gram method, considered to be a highly effective method, and how it is combined with a refinement, namely removal of content words, that has been used with other methods but not with the binary n-gram method.

1.1 Authorship Attribution

Authorship attribution is an old problem going at least as far back as the Old Testament where the Gileadites tested a man's pronunciation of "Shibboleth" to determine if he was from the opposing Tribe of Ephraim [1]. More contemporary examples of authorship attribution include the authorship of a number of works during Shakespeare's era, including the anonymous The True Tragedy of Richard III [2], and even works allegedly written by Shakespeare, such as The Two Noble Kinsmen [3], [4]. Another well-known example is the Federalist Papers, written between 1787 to 1788, a corpus of 85 pseudonymous articles in favour of ratifying the US Constitution and separating from Britain [5]. The papers were written by three authors all under the single pseudonym, "Publius", with the authorship of each being a closely-guarded secret. Years after their publication the authors of the Federalist papers came forward and disclosed the authorship of a number, although not all, of the papers. This made the Federalist papers one of the most-studied authorship attribution problems, having a known set of candidate authors, a known set of

attributed texts each written by a single author, and a set of disputed papers, also known to have been written by a single one of the candidate authors [6].

While authorship attribution is typically defined as determining who wrote what, it is more broadly defined as “any attempt to infer the characteristics of the creator of a piece of linguistic data” [7]. Although authorship attribution can include other media, text is the focus of the majority of authorship attribution research [7].

1.2 Using n-gram Analysis for Authorship Attribution

One of the most effective authorship attribution methods is the use of character n-grams to seek out frequently-occurring letter combinations in text. Historically, such analysis was used to break encryption methods using substitution of letters to make statistically-informed guesses about the probable plain text; for example that if the letter e is the most common letter in English, and the letter q has the same relative frequency in the ciphertext, then there is a good chance that e has been encrypted as q in the ciphertext. This frequency analysis cryptanalysis method is so successful that it motivated the development of block ciphers. Also, Gaines documents the most common 2- and 3-character grams in different languages, showing evidence that such n-grams are distinctive markers of languages [8].

The same analysis of relative frequencies has also been used to determine the authorship of documents. The premise is that individual writing styles arise from more regular use of certain words or phrases, and this would reflect in higher relative frequencies of those words in the author's text. The same observation was extended to the use of not words, but characters as the unit of analysis, partly on the basis that character n-grams are not delimited by whitespace and hence they can capture whitespace in their analysis, allowing frequencies of adjacent words to be represented but without representing the full word pair. It also fixed the size of grams, which cannot be predicted or limited in analysis of word grams.

Kestemont [9] states that "character n-grams have proven to be the best performing feature type in state-of-the-art authorship attribution". Stamatatos [10] found that an n-gram length of 4 characters appeared to be ideal for the English language, although other languages needed longer n-grams. This observation held true even when assessing short texts such as tweets [11].

Peng et al. [12] introduced the use of n-gram analysis based on a binary representation of text, and applied it to authorship attribution over social media postings. It takes the reduction of the unit of analysis a step further, from character to bit level and was motivated by binary having fewer options for each unit (two) than characters and hence was primarily a performance measure.

In this paper, the use of binary n-grams instead of character or word n-grams was motivated by two observations.

Firstly, they yield more frequencies for a smaller quantity of text. Peng's results [13] showed reasonable attribution accuracy with n-gram sizes of up to 16 bits, much smaller than the character 4-grams that Rocha et al. [11] found was the best performing n-gram size. This may be because the number of frequencies is maximised by reducing the text into the smallest possible components, i.e. bits. So when operating a sliding window sampling over binary data, there are many more different frequencies that occur across the same number of characters. It is also possible that the apparent improvement by binary methods over character methods could be attributed to the size and source of the texts being analysed, as the input text was even smaller for the character 4-gram experiments, dealing with tweets. The binary n-gram experiments were evaluated over a range of text sizes, but generally over larger ones. Because no direct comparison has been made between the binary and character n-gram methods, section 2 of this paper does so.

Secondly, binary n-gram analysis operates independently of either language or character encoding. While Kestemont [9] points out that character n-grams over multiple languages work well, the languages reviewed there all have the same Roman character set. While including the small number of characters in Cyrillic and Greek alphabets is still feasible (although coming with a performance penalty), it becomes more challenging when using larger alphabets such as in Asian pictograms, including in mixed-language texts.

Binary n-grams seem to have been little used for any purpose, with one exception being for graphing the output of encryption algorithms [14], so they have been hitherto under-utilised. Since Peng [13] found they were potentially beneficial for authorship attribution, further evaluation is justified.

In section 2, binary n-grams were compared directly to character n-grams and bag of words. Character n-grams are the nearest form to binary n-grams, using very similar methods, i.e. tallying the occurrences of specific n-grams using a sliding window approach to sample the text. Also, character n-grams are claimed to be "the most successful feature in both single-domain and cross-domain Authorship Attribution", and shown to be effective primarily because they capture information about affixes and punctuation [15]. Since Rocha et al. [11] identified character 4-grams as the best length (albeit on tweets), 4-grams and 5-grams were used for the comparison.

1.3 The Effect of Stop Words and Function Words on Authorship Attribution

While Peng's prior work established the authorship attribution potential of n-gram analysis, it also uncovered a number of limitations. One such limitation occurred when analysing a collection of 23 research writings which included a number of multiple-authored papers over two research areas [13]. This meant that any characteristic writing style of individuals was partly obfuscated by the presence of other authors on the same paper, so that other content-based similarities between papers, in particular the language specific to the research area, dominated the attribution, resulting on a higher rate of false attributions.

This paper addresses that limitation by determining how much influence content words have in a similarity assessment. Authorship attribution assesses similarity between documents based on the authors' writing styles, which is in principle the same process as information retrieval which assesses similarity between documents based on their content. Information retrieval measures similarity based on what the documents are "about", disregarding the authorship, analysing the content words that appear in documents while discarding words that do not contribute much to the document's meaning. So it is plausible that if content words are not purged from the text being analysed, the content words will increase the apparent similarity of documents, compromising the accuracy of applications that seek to assess similarity of documents for different purposes where the content is not as important, in particular, for authorship attribution.

A number of works assess the value of removing content words from authorship attribution analyses for these reasons. For example, Sundararajan and Woodard [16] found that "proper nouns are heavily influenced by content and cross-domain attribution will benefit from completely masking them". Sapkota et al. [15] combine character n-gram analysis with content word removal, although they focus less on removing whole content words than on ignoring less-productive parts of words. They consider character n-grams that include affixes, namely the beginnings (prefixes) and endings (suffixes) of words, along with white space prefixes and suffixes. They also include n-grams that feature punctuation at either end or within. They claim that affixes and punctuation "account for almost all of the power of character n-grams as features". They note however that the language studied (English) is more amenable to prefix n-gram analysis than suffix, and that because it has little inflection, the findings for English may not be repeated for other languages. It is also possible that some forms of social media content such as tweets may contain proportionately less of these important part-words, and may be less amenable to this form of analysis.

1.4 The Contributions of This Paper

This sequence of experiments investigates further aspects of binary n-gram authorship attribution. It starts by narrowing the scope of the text corpus to a closed set of well-known texts, mostly whole novels sole-authored by well-known individuals, with exceptions limited to a few, deliberate, instances incorporated during later experiments. As Peng et al. [17] note, aside from control accounts created for their experiment, "the collected data cannot enable validation of the results against known authorship". This is because their text corpus was compiled using a web scraper on two online discussion fora, where the author identities of the texts is unknown and any relationships between authors is what is being detected (e.g. astroturfers).

This paper combines the binary n-gram method with the purging of content words from the text prior to analysis. A sequence of controlled experiments assesses the influence of content words on the accuracy of binary n-gram analysis, finding that in some contexts, the binary n-gram method applied over text with content words removed offers a competitive or better authorship attribution method. The experiment sequence is intended to isolate the effect of function and stop words on binary n-gram authorship analysis in a range of contexts.

The paper starts with the motivation for the problem and then reviews of authorship attribution (Section 1). The experiments are reported, starting with the accuracy of binary n-gram analysis as an authorship attribution tool (Section 2). The influence of different values of n (i.e. different n-gram sizes) is explored, and establishes that n=14 appears to be a reliable test value for subsequent experiments (Section 3). The removal of content words is the next step, and an experiment then analyses the same texts as in the previous experiment but this time with only stop words remaining (Section 4). A larger set of works was then assessed, with works from the same fictional universes but different authors, as well as cross-genre works (Section 5). The paper closes with a discussion (Section 6) and finally conclusions and future work (Section 7). Methodologies for each experiment are included in the relevant section.

2. COMPARING BINARY N-GRAMS WITH CHARACTER N-GRAMS

This experiment aims to demonstrate whether a binary n-gram approach is at least as effective at authorship attribution as other common methods, such a character n-grams and bag of words, in at least some contexts. This section's results also form a baseline of performance, for comparisons with the subsequent experiments.

This first test of authorship attribution evaluates binary n-grams on the Federalist papers that were the foundation of the United States of America's independence, examined in the Introduction.

2.1 Methodology

This experiment uses a corpus of 24 texts from the Federalist journal to compare the performance of binary n-grams to other methods of authorship attribution. Two texts were used as training data for the system with each containing concatenated text from either Hamilton or Madison but not both. Text file "H" contained 6 concatenated papers known to be from the author Hamilton, while text file "M" contains 6 concatenated papers known to be from the author Madison. The remaining 12 text files were used as test data by the system using either binary n-grams, character n-grams, or bag of words and compared against the training data in order to classify the text as either written by Hamilton or Madison. The 12 selected test texts were all single papers published in the Federalist, with known authors, and contained 6 known to be written by Hamilton, and 6 known to be written by Madison.

For binary n-grams, all standard text was processed in place into 7-bit ASCII binary. Non-standard text, which could not be represented by ASCII was processed and placed into 16-bit UCS2. Once all text was processed into binary, the frequencies of all binary n-grams were recorded in an array using the sliding window technique, with the sliding window set to 14 bits in

length, i.e., n=14. The array was in a fixed order, with the element of the array directly matching the binary n-gram.

For the character n-gram method, no processing took place, and the frequencies were recorded into a Dictionary<String,Int>, using the same sliding window technique, with the sliding window set to 5 characters in length. The Dictionary key was the n-gram character combination, whilst the Int was the frequency. The Dictionary was then sorted alphabetically.

For the bag of words method, no processing took place, and the frequency of each whole word was recorded into a Dictionary<String,Int>, with the String, key, being each unique word, and the Int, the frequency. For the bag of words method, each single whole word was considered, and all punctuation was removed and was not included in the Dictionary data structure. The Dictionary was then sorted alphabetically.

For the measurement of similarity between texts, a cosine distance was calculated between the training text ("H" or "M") and each of the 12 validation texts. The cosine distance was calculated by treating the list frequencies found for the observed patterns as a vector, with each item in the frequency list contributing to the vector's dimensions. As a simple example, Table 1 shows 2 texts analysed by a binary n-gram approach.

N-gram	Frequency Text A	Frequency Text B
00	10	5
01	5	10
10	2	10
11	0	2

TABLE 1: Frequency table for simple 2-gram binary example.

In this example Text A would then be represented as the vector [10, 5, 2, 0], and Text B would be represented as the vector [5, 10, 10, 2].

Once the vectors had been prepared, a cosine distance measurement between the training and validation texts and their corresponding vectors was then performed.

To compare the vectors, the number of dimensions for each vector needed to be identical. As the array of binary n-grams were always of a fixed length, 2n, no processing on this array was required to calculate the cosine distance, as each vector generated from the array had the same dimensionality.

For the bag of words, and character n-gram methods, since the Dictionaries only contained observed patterns - the dictionaries to be compared were first processed to add all the patterns found in Dictionary A, and not in Dictionary B, to Dictionary B with 0 frequency, and vice versa, to ensure the dimensionality of the resulting vector was identical. The Dictionaries were then sorted alphabetically once more.

To classify each of the validation texts - the distance to training text "M" was subtracted from the distance to training text "H", which are both on a 0 to 1 scale (with 1 indicating the texts were identical). If the result of the subtraction remained positive, the text was classified as written by

Hamilton, and if the result was negative, the text was classified as written by Madison. This usage of cosine similarity is based upon, and nearly identical to, a similar experiment carried out on the Federalist papers using a character n-gram analysis [18].

Finally, each method was scored based on how many texts it classified correctly.

2.2 Results and Discussion

The experiment showed that binary n-grams were at least as effective as bag of words, and slightly more effective than character n-grams.

Table 2 shows that all methods achieved roughly the same attribution percentage, with the bag of words and the binary methods correctly classifying an additional text.

Method	Correct Attribution
Character N-grams (n = 5)	83.33%
Bag of Words (Single Word)	91.66%
Binary N-grams (n = 14)	91.66%

TABLE 2: Attribution score table for binary n-gram and other methods.

A caveat with the results is that the classification is based on a limited number of samples. This is due to Madison only being officially recognized as the sole author of 14 of the papers [19]. It is preferable to have an equal number of Madison and Hamilton papers, so that an incorrectly biased result, for example, that all papers were attributed to Hamilton (when Hamilton comprised of 80% of the texts), would not provide results that appeared favourable. Further, we wanted to use at least half of the available texts as training. These factors provided a strict cap on the number of papers available to test the system with. Further, as a weakness of supervised learning in general, many of the classified instances demonstrated extremely weak attribution but had to be classified anyway. Indeed, one paper in particular was correctly classified as a Hamilton paper on a very small margin of 0.000852023. While this was technically above zero,and hence classified as being written by Hamilton, as one might imagine from simply seeing how many leading zeroes the number has, this was hardly a definitive result, and indeed the classification often swapped between Hamilton and Madison when changes were made to classification algorithm, such as increasing or reducing the n-gram size or changing papers in the training corpus.

On the topic of the selected size of n-grams, a later experiment tests the variation and impact that differing binary n-gram sizes has on the data, and find that 14-grams provide a good trade-off between computational complexity and quality of results. For the character n-gram, the usage of 5 n-grams was based on others' research that character n-gram were effective from the n=4, to the n=10 ranges [20], and used a number within this range whilst keeping the processing time manageable.

2.3 Section Summary

As demonstrated above, binary n-grams can perform at least as well as other, similar, methods such as character n-grams and bag of words which are commonly used for authorship attribution. Binary n-grams achieved a 91% accuracy score for identifying the author, identical to bag of words, and slightly better than character n-grams which achieved an 83% accuracy score. Therefore, binary n-grams are a relevant avenue for further inquiry. Further, the accuracy of attribution (91%) for binary n-grams is slightly higher than similar research carried out by Kjell,

Woods & Frieder [18], which also used cosine distance as a measurement tool and had a very similar methodology, and reported an accuracy of 89.2% using two-character n-grams, and 87.6% accuracy using three-character n-grams. However, the work here only examines a very small subset of texts, all on a similar topic (i.e. a single-domain case), which only needed to be classified between two known authors, where all texts are believed to be attributable to those two authors. To show that binary n-grams retain their authorship attribution effectiveness more generally, increasing the number of authors and number of texts, an experiment on this topic is documented in section 5.

3. EFFECTS OF N-GRAM SIZES ON BINARY N-GRAM AUTHORSHIP ATTRIBUTION

The previous section showed that in some contexts, using n-grams was as least as effective as using other methods of authorship attribution, such as character n-grams and unigram bag of words. The purpose of this section's experiment is to perform some objective measures of how binary n-gram size affects authorship attribution properties. Also, the previous experiment only classified texts into two authors. The new experiment increased the number of authors to 5 to assess how well the binary n-gram approach scales as the number of authors increases.

The selection of texts is for 3 fictional works from each of 5 authors. In most cases, the author's 3 works are from the same fictional universe (Tom Clancy, Dorothy Penrose, Sir Arthur Conan Doyle), implying some sharing of proper names. However one author's works (James Joyce) were from different fictional universes, while the last author (J.K. Rowling) had two texts from one universe while the third was from a different universe.

3.1 Methodology

This method uses the k nearest neighbours (kNN) method to decide which of the available texts is the closest to the one being analysed.

15 novel-length texts were selected from 5 well-known authors.

Each of the texts was analysed using 7-bit encoding, and converted to binary strings.

Each text was then analysed using (initially) 10 bit n-grams, generating a list of each n-gram frequency.

Each text was then cross-compared to the other 14 texts by using the cosine distance measure described in the previous section.

The cosine distance between each text was recorded in a table.

The above steps were repeated for all n-gram sizes 10 to 18 inclusive.

Once all processing had taken place, the results were recorded in a table.

Once the results had been recorded for each text, the cosine distance scores were ranked from high to low (most similar to least similar).

Once ranked, the top two results (those most similar) were selected as the text’s k nearest neighbours (i.e. set k = 2).

For each of the 2 nearest neighbours chosen, a score equal to the cosine similarity was awarded if the correct author was chosen as a neighbour. An incorrect selection was penalized with negative points also equal to the cosine similarity.

The 3 texts for each author, and their 2NN scores, were averaged, giving the author a score from -1.0 to 1.0, with -1.0 indicating that no correct authors were identified, and that the texts were completely dissimilar to each other, while 0 indicating the correct author was identified 50% of the time, and 1 indicating that all texts were attributed to their correct author, and further, that the texts were exactly similar (i.e. identical).

Finally, the results for each author were calculated and listed in a table.

3.2 Results and Discussion

The results are summarized in Table 3.

n-gram size	Penrose	J.K. Rowling	Sir Arthur	James Joyce	Tom Clancy	Total Score
10	100.00%	33.33%	83.33%	33.33%	0.00%	50.00%
11	100.00%	100.00%	83.33%	50.00%	0.00%	66.67%
12	100.00%	100.00%	83.33%	66.67%	33.33%	76.67%
13	100.00%	66.67%	83.33%	66.67%	66.67%	76.67%
14	100.00%	66.67%	83.33%	66.67%	100.00%	83.33%
15	100.00%	50.00%	83.33%	66.67%	100.00%	80.00%
16	100.00%	33.33%	83.33%	83.33%	100.00%	80.00%
17	100.00%	33.33%	83.33%	83.33%	100.00%	80.00%
18	100.00%	33.33%	83.33%	83.33%	100.00%	80.00%

TABLE 3: Percentage attribution scores for binary n-grams n=10 to n=18.

Table 3 shows that n-gram size does make a significant difference to the author attribution score. However, some authors' attributions were considerably more static than others. This can especially be seen by the author Penrose, whose texts attributed correctly, regardless of n-gram size. By contrast, J.K Rowling was the least consistent, with the 11- and 12 n-grams correctly attributing all of her texts, whilst the 18 n-gram was only able to attribute 33% correctly. In general, though, the classification accuracy increased as the n-gram size increased until a plateau was reached. The most interesting feature of this data, however, is that n = 14 achieved the highest total score, and attributed 1 more text correctly than the next highest.

In order to understand this n = 14 feature, and to better understand the results in general, it is important to understand how kNN works, and why it might give us the results shown.

With kNN for k=2, the closest 2 neighbours (where closeness is measured with cosine similarity) are discovered. In traditional kNN classification, the sample is automatically classified as the type with the highest count of samples in the k closest samples.. Since there are more triangles, the sample would be classified as a red triangle.

However, a weakness in the 2-author experiment in section 2 is that the sample MUST be classified regardless of how small the difference between the data is, and how close the samples are together. Very poor classification would be achieved if all the samples were very spread out or very close, since the k nearest neighbours, and only those k neighbours, always decide the classification of the sample regardless of how close or far away the k neighbours are. Changing the value of k could change the classification of the sample.

This is relevant to this experiment because of just how close many of the neighbours were. A full listing of all the cosine similarities is available on request, but as a few examples, for n = 14, J.K. Rowling's books were correctly classified on a minuscule margin of just 0.0013945. To further provide context, the entire range of cosine similarity for J. K. Rowling's texts was between 0.9377 and 0.9980, with 1.0 being identical. As this classification based on such a small margin, it is possible that n = 14 having a higher classification accuracy is an artefact of this similarity measure, and that the general trend of higher n-grams providing better classification might otherwise be observed.

Another interesting observation regarding this data is that books written by the same author using the same characters and setting generally had much better classification accuracy than books using different settings or characters. For example, Penrose was the most successfully-classified author, but, taking a closer look at the texts used to classify this author, specifically, *Dorothy Dale's Promise*, *Dorothy Dale in the West*, and *Dorothy Dale and Her Chums*, it is evident that every one of these books is part of Margaret Penrose's "Girls Series" books, based around the same main characters, and one might conjecture that from a simple word or character perspective, the word "Dorothy" appearing several hundred times per book would make the texts more similar to each other.

This conjecture is supported by looking at cosine similarities from J.K Rowling's work, including *Harry Potter 2*, *Harry Potter 3*, and also *The Casual Vacancy*. Unsurprisingly, the two Harry Potter books always correlate extremely strongly, since they share character and place names (such as "Harry" and "Hogwarts", etc.), and are, in general, about the same broad topic of witchcraft and wizardry. *The Casual Vacancy*, on the other hand, is very different, intended for an adult audience, is set in Pagford rather than Hogwarts, and features themes of politics, drugs, and prostitution that were absent in Harry Potter. It is little wonder, then, that this particular text is much harder to associate with the other works by J.K Rowling, because the absence of shared content words is going to drastically reduce their similarity. And, while we are specifically looking for characteristics that make each author unique, elements such as character names are not always going to be useful when considering the wide variety of texts and contexts that a single author can produce, especially when fan fiction is considered.

3.3 Section Summary

The conclusions from this experiment are that, in general, increasing the size of the n-gram increases the authorship classification accuracy to a reasonably accurate 80% (vs random chance of 20%), which then plateaued until the experiment concluded. However, a principal finding of this experiment is that content words, such as character names and places, do provide a large influence on whether the text will be successfully attributed or not, and can hamper attribution to an author when the topics are different enough.

This is undesirable, as not only does this decrease the effectiveness of the authorship attribution of the selected novels, especially when a single author has written multiple books in different fictional universes, such as J.K Rowling's Harry Potter series, compared to another book, *The Casual Vacancy*, but also in a broader sense, topic sensitivity would drastically hinder the aim of authorship attribution on social media. A single author on social media is likely to be writing about any number of things, from politics to Prada. Therefore, this experiment will be repeated, but with all of the content words removed from the text during the pre-processing stage. Doing so should increase the effectiveness of the authorship attribution by only leaving "style" words than are indicative of the way an author writes, rather than basing the authorship attribution predominantly

on what they are writing about. The next section performs the same experiment on the same texts, except for the removal of the content words during the pre-processing stage.

4. EFFECTS OF STOP WORD REMOVAL ON BINARY N-GRAM AUTHORSHIP ATTRIBUTION

This experiment seeks to test the effect of filtering out content words on the accuracy of the binary n-gram authorship attribution method. The section will repeat the previous experiment, but with all content words removed from the texts, to test whether content words, and content sensitivity are tainting authorship attribution. It provides a direct comparison for the same texts and same methods, isolating the removal of content words as the variable of the experiment.

4.1 Methodology

The same 15 novel-length texts from the previous experiment were selected.

Each of the texts was pre-processed and analysed using a publicly-available stop-word [22].

The pre-processor removed all words in the text which did not appear in the stop word list, leaving the remaining words in the same order as before.

Each of the filtered texts was analysed using 7-bit encoding, and converted to binary strings.

Each filtered text was then analysed using 10-bit n-grams, generating a list of each n-gram frequency.

Each text was then cross compared to the other 14 texts by using the cosine similarity measure.

The cosine similarity between each text was recorded in a table.

The above steps were repeated for n-gram sizes 10 to 18 inclusive.

Once all processing had taken place, final total and aggregate results were recorded.

Once the results had been recorded, for each text, the cosine distance scores were ranked from high to low (most similar to least similar).

The top two results (the two most similar) were selected as the text's k nearest neighbours (k = 2).

For each of the 2 nearest neighbours chosen, a score equal to the cosine similarity was awarded if the correct author was chosen as a neighbour. An incorrect selection was penalized with negative points also equal to the cosine similarity.

The 3 texts for each author, and their kNN scores, were summed together and averaged, giving the author a score from -1.0 to 1.0 inclusive, with a -1.0 indicating that no correct authors were identified, and further, that the texts were completely dissimilar to each other, a 0 indicating the correct author was identified 50% of the time, and a 1 indicating that all texts were attributed to their correct author, and further, that the texts were exactly similar.

The results for each author were calculated and listed in a table, and these results for stop words only were compared to the results of the previous experiment for all text.

4.2 Results and Discussion

The results for this experiment are given in Table 4.

n-gram size	Penrose	J.K. Rowling	Sir Arthur	James Joyce	Tom Clancy	Total Score
10	100.00%	33.33%	83.33%	100.00%	66.67%	76.67%
11	100.00%	33.33%	83.33%	100.00%	100.00%	83.33%
12	100.00%	33.33%	83.33%	100.00%	100.00%	83.33%
13	100.00%	33.33%	83.33%	100.00%	100.00%	83.33%
14	100.00%	33.33%	83.33%	100.00%	100.00%	83.33%
15	100.00%	33.33%	83.33%	100.00%	100.00%	83.33%
16	100.00%	33.33%	83.33%	100.00%	100.00%	83.33%
17	100.00%	33.33%	83.33%	100.00%	100.00%	83.33%
18	100.00%	33.33%	83.33%	100.00%	100.00%	83.33%

TABLE 4: Percentage attribution scores for binary n-grams n=10 to n=18.

As can be seen from Table 4, the stop words-only analysis gave a modest increase in authorship attribution effectiveness. Specifically using stop words only increased the plateau from 80% correct attribution to 83.33%, an increase of one extra successfully-attributed text. This does disregard the single peak 83.33% attributions score in the previous experiment at n = 14, although that score may have been an not representative of the general trend displayed by the other results.

Further, as can be seen from the graph, the peak effectiveness of authorship attribution for stop words was reached at n = 11, much less than the n = 14 of the previous experiment. Additionally, there were no outlier or unusual results recorded in this experiment, unlike the n = 14 outlier of the previous experiment.

Looking more carefully at the data, we discuss exactly for which authors this method of removing content words improved the authorship attribution, as well as, perhaps more importantly, for which authors it did not. Authorship attribution for James Joyce was the main success of removal of content words. In the previous experiment, which included full text, the *Dubliners* text was attributed most closely to *Portrait of a Young Man* (written by Joyce) and *A Study In Scarlet* (written by Sir Arthur Conan Doyle), but when considering only stop words, the *Dubliners* text was attributed most closely to *Portrait of a Young Man* and *Ulysses*, both written by Joyce. This is an encouraging result, because it demonstrates the apparent strengths of this method – James Joyce has a quite unique writing style, but the *Dubliners* text is a collection of short stories by Joyce, and features many different character names and other content sensitive content, whilst *Ulysses* is a single story and, again, features different characters to other works by Joyce.

However, whilst focusing on the stop words for Joyce improved authorship attribution, it did not improve authorship attribution for J.K. Rowling. It might be assumed that removing the topic would assist in the author attribution between the Harry Potter books, which otherwise attribute well to other Harry Potter books, but very poorly to Rowling's *The Casual Vacancy*. However this failed to happen.

There are many possible reasons for this. Perhaps J.K Rowling's style changed, as she targeted a different audience, or as she developed as an author. Further, J.K. Rowling was writing other novels during this time as Robert Galbraith [21], and may have intentionally disguised her style, and that some of this deliberate style change "rubbed off" on her writing of *The Casual Vacancy*. Another possibility is that removing the content words is sound in principle, but that the selected stop words list is not "smart" enough to accurately capture the style of the author. Using full text for *The Casual Vacancy*, the Harry Potter books were the 5th and 6th closest neighbours, with scores of 0.9479, and 0.9477 respectively, compared to the 0.9634 of the closest neighbour, a Tom Clancy book. However, with the stop words only, the Harry Potter books moved into 2nd and 4th spot, with scores of 0.9544 and 0.9521, which was an improvement, but still not enough to displace the same Tom Clancy book from 1st place.

Looking at the list of stop words [22] may provide some insight. For example, it includes terms like "he", "she", "I", etc. However, if the author writes two books, one about Alice, and one about Bob, a story about Alice is likely to use pronouns such as "she", "her", etc., whilst a story about Bob is going to use "he" and "him". Also, a story in the first person is going to make more frequent use of "I" than a story in the third person. So, whilst the stop word list was expected to detect the style of the author, content-based differences to the story such as character gender would still affect the authorship attribution. As a complicating counterpoint, Tolkien's works are famous for only including "she" 12 times during the entire Lord of the Rings Trilogy, and this choice of a predominantly male cast could reflect his style, and would be assessable against his other works, such as the similarly-masculine The Hobbit [23]. Furthermore, as stop word lists are used by search engines to filter out common, non-useful words from search strings [22], it may be that they are unsuitable for authorship attribution, and that a "smarter" list must be developed to capture "style words" indicative of the style of an author (also termed *cinnamon words*) [24], rather than purely functional words devoid of meaning. An author may also favour classes of words, for example, Tolkien uses archaic words such as *sward*, *hythe*, and *glede*.

Finally, removing all occurrences of words other than the 300 or so contained in the stop word list used dramatically decreases the total amount of text to work with. Whilst this was not an issue in the above experiment, which used feature-length texts with tens of thousands of words, if the goal is to apply this to a social media context, where text length is reduced, especially in character-limited social media sites like Twitter, it could seriously impact the applicability of analyses which exclude any text for whatever reason.

4.3 Section Summary

This experiment found a modest improvement of 3.33% in the effectiveness of author attribution by removing the content words from the text. This leads to the conclusion that removing content words is a useful technique, since it performs no worse than the full text experiment, and provides a modest increase where the author writes books using separate characters and locations. However, the increase was not as large as was desired, and while it succeeded at increasing the correct attribution for James Joyce, it failed to improve attribution for J.K. Rowling, whose two very different fictional universes, Harry Potter, and *The Casual Vacancy*, provided the original motivation for conducting this content-insensitive experiment.

Because stop word lists contain words only function words, which include articles such as "he", "she", they are still heavily influenced by the topic the author is writing about, for example, a male or female character. Therefore, future work needs to be done to ascertain a list of appropriate "style words" which are more indicative of the style of the author and less reflective of the content of their works.

Finally, pre-processing the text to leave only stop words significantly reduces the amount of text to run the binary n-gram test on. Whilst this was not an issue with the current corpus, owing to the significant length of each of the novels it contains, if the ultimate goal as to apply this technique to social media, this needs to be carefully monitored, as a poorly-selected stop word list may reduce the quantity of available text to process for smaller length social media, such as Twitter, to a few, or even zero, remaining words.

5. AN UNSUPERVISED EXAMINATION OF TEXTS

To further test the binary n-gram method, two further aspects were introduced into the experiment, and a third aspect is investigated more extensively.

The first aspect is to examine a text that shares a topic with other texts, but is constructed differently, i.e. a cross-genre scenario. The chosen text for this was the latest Harry Potter story *The Cursed Child* and compared it to other Harry Potter books. *The Cursed Child* is unique amongst the Harry Potter fiction, is that it is a published stage play [25], rather than a novel like the other Harry Potter books. This text shares the Harry Potter world, but is very different in its execution – for example, a scene in *The Cursed Child* may look something like this:

> “HARRY points out RON, HERMIONE, and their daughter, ROSE. LILY runs hard up to them.
> Uncle Ron. Uncle Ron!!!” [25]

By contrast, a scene from the other Harry Potter books is more likely to look something like this:

> “Neville hung his head. Professor McGonagall peered at him through her square spectacles.
>
> "Why do you want to continue with Transfiguration, anyway? I've never had the impression that you particularly enjoyed it.” [26]

The Cursed Child text uses capital letters for names, omits quotes for speech, uses less detailed descriptions of character actions, amongst other literary changes. Sundararajan and Woodard [16] define "Cross-domain scenarios include both cross-topic (same genre but different topics) and cross-genre (different genres) scenarios" so the previous experiments have focused on single-topic and cross-topic works, while this experiment additionally incorporates cross-genre works, and examines how these differences affect the authorship attribution for this latest Harry Potter book.

In addition to just Harry Potter, we also wished to test more broadly the pseudonym of J.K. Rowling, Robert Galbraith. When we examined how J.K. Rowling’s authorship attribution of Harry Potter books compared to *The Casual Vacancy* text in the previous experiments, they were inconclusive in that only a single non-Harry Potter book was amongst the corpus tested. This next experiment, however, includes all the Harry Potter books, as well as *The Casual Vacancy*, plus three additional books J.K. Rowling has published under the Robert Galbraith pseudonym. We mentioned in the previous experiment that it seems that topic-sensitivity was a major limitation of full-text text analysis, so these additional J.K. Rowling texts have been included so as to draw a more robust conclusion regarding to what extent the same author writing about a different topic can affect authorship attribution, making this a cross-topic test.

The second new aspect is to investigate multiple-authored works, so this section considers the authorship attribution for Tom Clancy novels. Tom Clancy was the sole author of many of the "Jack Ryan" universe novels, however, as Tom Clancy’s age advanced, and the series progressed, Clancy collaborated with Mark Greaney on a number of the Jack Ryan books [27], and, after Tom Clancy died, Mark Greaney continued publishing Jack Ryan books under the Tom Clancy name, along with other authors, such as Grant Blackwood. Therefore, this experiment

aims to detect whether a Tom Clancy text was written by Clancy himself, written in collaboration with another author, or written by another author entirely. In a sense, this part of the experiment is testing the opposite of the J.K. Rowling experiment, which seeks to detect a single author over multiple topics, whereas this Tom Clancy experiment looks to detect multiple authors over a single topic.

Finally, a number of texts are included that have no relation to either J.K. Rowling or Tom Clancy. This included novels written by J.R.R Tolkien, as well as academic papers and academic works written by researchers at UniSA. These unrelated texts were important for testing the strength of authorship attribution when attempting to differentiate between related texts and "noise".

5.1 Methodology

42 texts from a variety of authors were prepared from a variety of sources including novels, academic papers, and lecture notes. This included 20 texts from the Tom Clancy universe, written by either Tom Clancy, co-authored between Tom Clancy and Mark Greaney, or written entirely by Mark Greaney or Grant Blackwood.

It also included 12 texts by J.K. Rowling, these being texts written about the Harry Potter universe, J.K. Rowling's writings under the Robert Galbraith pseudonym, as well as texts under the J.K. Rowling name not in the Harry Potter universe.

The cosine distance for full text, stop words, stop words without spaces, and content words without spaces between the J.K. Rowling and Tom Clancy texts was calculated at n=7, using 7-bit standard ASCII, and compared to every other text.

The results were recorded, with any interesting observations likewise noted.

5.2 Results and Discussion

The full results of this experiment are available on request, but this section includes results of most interest.

In this experiment, one of the things we were looking for was how context affects authorship attribution, i.e. what form the text takes, such as comparing a novel written by J.K. Rowling, and comparing it to a screen play J.K. Rowling collaborated on. Again, the topic of Harry Potter remains identical, yet the text is portrayed differently. A truncated set of results for *The Cursed Child* is in Table 5.

Harry Potter Cursed Child - JK Rowling Ref text	**Cosine Similarity**
Harry Potter 1 - JK Rowling	0.890144154
Tom Clancy Under Fire - Grant Blackwood	0.885800401
Harry Potter 3 - JK Rowling	0.883005526
Harry Potter 2 - JK Rowling	0.882128003
Harry Potter 5 - JK Rowling	0.876933585
Harry Potter 7 - JK Rowling	0.875973105

Dead or Alive - Tom Clancy Grant Blackwood	0.875396414
The Silkworm - Robert Galbraith	0.87411465
Harry Potter 6 - JK Rowling	0.873821221
Harry Potter 4 - JK Rowling	0.873588089
…	…
MD01 ht2016 - Helen	0.710436731

TABLE 5: Cosine similarity scores for reference text "The Cursed Child" using full text comparison.

Table 5 displays the top 10 most similar results, as well as the least similar result. The results are encouraging, since this text is correctly grouped with other J.K. Rowling texts from the Harry Potter series, and even manages to identify a text written by Robert Galbraith (J.K Rowling's pseudonym). We also examined the stop word similarity in Table 6.

Harry Potter Cursed Child - JK Rowling Ref text	**Stop word Cosine Similarity**
Harry Potter 1 - JK Rowling	0.87494083
Tom Clancy Under Fire - Grant Blackwood	0.872042206
Harry Potter 3 - JK Rowling	0.870705941
Harry Potter 2 - JK Rowling	0.869252184
The Silkworm - Robert Galbraith	0.854853828
Harry Potter 5 - JK Rowling	0.849511343
Harry Potter 6 - JK Rowling	0.844170792
Harry Potter 4 - JK Rowling	0.836333339
Harry Potter 7 - JK Rowling	0.834615164
Dead or Alive - Tom Clancy Grant Blackwood	0.823979339
…	…
MD04	0.568677587

TABLE 6: Cosine similarity scores for reference text "The Cursed Child" using stop word only comparison.

Again, this result looks encouraging. While a few texts have changed positions, the level of attribution is still quite good, with 8 of the top 10 texts being correctly attributed to J.K. Rowling.

However, we should also look at the converse of this – if *The Cursed Child* attributes well to Harry Potter texts, do the Harry Potter texts attribute well to *The Cursed Child*? This poses an interesting question because *The Cursed Child* was not written entirely by J.K. Rowling, but rather was a collaborative effort between Jack Thorne, John Tiffany, and J.K. Rowling, with Thorne being directly responsible for writing the script. To conduct this comparison of J.K. Rowling texts, we selected the text with the highest attribution for both full text and stop words only, Harry Potter 1. An excerpt of the results are in Table 7.

Harry Potter 1 - JK Rowling Ref text	**Cosine Similarity**
Harry Potter 2 - JK Rowling	0.995397803
Harry Potter 3 - JK Rowling	0.994741247
Harry Potter 4 - JK Rowling	0.993617601
Harry Potter 5 - JK Rowling	0.992726032
Harry Potter 7 - JK Rowling	0.992682162
Harry Potter 6 - JK Rowling	0.991573339
Dead or Alive - Tom Clancy Grant Blackwood	0.975760207
Without Remorse - Tom Clancy	0.974316001
Patriot Games - Tom Clancy	0.973491073
The Teeth of the Tiger - Tom Clancy	0.971688271
…	…
Harry Potter Cursed Child - JK Rowling	0.890144154
MD04	0.864306145
MD10	0.863962359
MD02 IJIM2014 - Helen	0.837490776
MD03 ICWL2009 - Helen	0.829197561
MD01 ht2016 - Helen	0.821026868

TABLE 7: Cosine similarity scores for reference text “Harry Potter and the Philosopher's Stone” using full text comparison.

And again for stop words only in Table 8.

Harry Potter 1 - JK Rowling Ref text	**Stop word Cosine Similarity**
Harry Potter 2 - JK Rowling	0.997836663
Harry Potter 3 - JK Rowling	0.99755717
Harry Potter 5 - JK Rowling	0.994414852
Harry Potter 4 - JK Rowling	0.993741677
Harry Potter 7 - JK Rowling	0.993141344
Harry Potter 6 - JK Rowling	0.992428518
Dead or Alive - Tom Clancy Grant Blackwood	0.987445695
Tom Clancy Under Fire - Grant Blackwood	0.986075112
The Teeth of the Tiger - Tom Clancy	0.98246844
Patriot Games - Tom Clancy	0.981961491
…	…
Harry Potter Cursed Child - JK Rowling	0.87494083
MD05	0.859898805
MD02 IJIM2014 - Helen	0.855835013
MD03 ICWL2009 - Helen	0.852940968
MD01 ht2016 - Helen	0.85253472
MD10	0.850955466
MD06	0.842016697
MD04	0.835403798

TABLE 8: Cosine similarity scores for reference text “Harry Potter and the Philosopher's Stone” using stop word only comparison.

Table 8 shows that the absolute attribution amount has not changed (remains at 0.890 similarity for all text, and 0.874 for stop words, as it should since the cosine measure is commutative). However, the ranking of *The Cursed Child* has dropped dramatically – instead of being a reciprocal number 1 attribution, *The Cursed Child* is now ranked 36th and 34th for full text and stop words only, respectively.

This result highlights the issues of context. The top result for Harry Potter 1 was Harry Potter 2, with a cosine similarity of 0.998, whilst the top result for *The Cursed Child*, which was Harry Potter 1, was just 0.875. This appears to be primarily because of the differences between writing a screen play and a novel. For example, a large dragon battle is likely to be absent in a screen play, because there are practical difficulties in performing such a scene for a live screen play, which do not exist in a novel. Further, descriptions of facial expressions and the like may be omitted from the screen play on the assumption that the audience either cannot see them, or the actor playing the character should develop their own take on scene. These differences in text resulting from differences in style extend beyond just a screen play vs novel example, too. For example, consider an academic paper vs a blog. Even if these two texts were on exactly the same subject, the tone and choice of words between these two texts is likely to be dissimilar. Indeed, these results indicate that an author's style changes depending on what context they are working in. For another relevant example that ties back to social media, consider the style change for an author between a blog and Twitter post. Since Twitter has a strict limit on the number of characters, an author may deliberately abbreviate, or omit words so that their message meets the requirements of Twitter, and further, the author may incorporate features like hashtags because these are available to them on Twitter, but the same features are not available to them on their blog.

Next, we examine how J.K. Rowling's pseudonym (Robert Galbraith) attributed to both the pseudonym and J.K. Rowling herself. We selected one of Robert Galbraith's books and the results are in Table 9:

Career of Evil - Robert Galbraith Ref text	**Cosine Similarity**
The Cuckoo's Calling - Robert Galbraith	0.986841847
The Silkworm - Robert Galbraith	0.986344526
The Casual Vacancy - JK Rowling	0.983995889
Without Remorse - Tom Clancy	0.978485921
Harry Potter 7 - JK Rowling	0.977513119
Tom Clancy Full Force and Effect - Mark Greaney	0.975106365
Harry Potter 5 - JK Rowling	0.975027506
Threat Vector - Tom Clancy with Mark Greaney	0.974705532
Dead or Alive - Tom Clancy Grant Blackwood	0.974240739

Patriot Games - Tom Clancy	0.973851802
…	…
Harry Potter Cursed Child - JK Rowling	0.85743321
MD02 IJIM2014 - Helen	0.854040917
MD03 ICWL2009 - Helen	0.847906131
MD01 ht2016 - Helen	0.83768359

TABLE 9: Cosine similarity scores for reference text "Career of Evil" using full text comparison.

And with stop words only shown in Table 10:

Career of Evil - Robert Galbraith Ref text	**Stop word Cosine Similarity**
Harry Potter 7 - JK Rowling	0.986569155
Harry Potter 5 - JK Rowling	0.985993908
The Casual Vacancy - JK Rowling	0.98507601
Harry Potter 4 - JK Rowling	0.98439207
Without Remorse - Tom Clancy	0.983759289
Harry Potter 6 - JK Rowling	0.983194627
Tom Clancy Full Force and Effect - Mark Greaney	0.982006415
Threat Vector - Tom Clancy with Mark Greaney	0.980197872
Tom Clancy Support and Defend - Mark Greaney	0.979118243
Harry Potter 2 - JK Rowling	0.978330858
…	…
MD03 ICWL2009 - Helen	0.869329281
MD04	0.866228014

MD06	0.860249001
Harry Potter Cursed Child - JK Rowling	0.821588339

TABLE 10: Cosine similarity scores for reference text "Career of Evil" using stop word only comparison.

This text was one of the most curious of those examined. Full text attributed well to both Galbraith (as expected with these novels sharing characters and places), and to J.K. Rowling's other texts, including both *The Casual Vacancy* and the Harry Potter series. There are however a number (5/10) of incorrect attributions.

However, once we switch from using full text to stop words, not only does the number of correct author attributions increase slightly for the top ten texts (6/10), the number of Harry Potter attributions increases from 2/10 to 5/10, but the number of attributed texts to Robert Galbraith drops from 3/10 to 0/10. It is hard to draw significant conclusions from this rather counterintuitive result, other than what we previously conjectured, that the stop word list is too naïve, and needs to focus more on style words. For reference, Robert Galbraith's other texts drop down to rank 12th and 18th, well below the Harry Potter texts. Further, the extremely poor attribution of *The Cursed Child* can also be seen here, where it achieves the single lowest attribution when comparing stop words only.

As a final test with J.K. Rowling, we repeated the attribution of J.K. Rowling's *A Casual Vacancy*, which was, again, a very different topic than her Harry Potter books, but was not written using a pseudonym, and compare it to a much larger variety of J.K. Rowling's texts, and observe how strong attribution was demonstrated. This can be seen in Table 11 and Table 12.

The Casual Vacancy - JK Rowling Ref text	**Cosine Similarity**
The Cuckoo's Calling - Robert Galbraith	0.989638842
The Silkworm - Robert Galbraith	0.984707666
Career of Evil - Robert Galbraith	0.983995889
Without Remorse - Tom Clancy	0.976923624
Harry Potter 7 - JK Rowling	0.975949303
Tom Clancy Full Force and Effect - Mark Greaney	0.975683828
Threat Vector - Tom Clancy with Mark Greaney	0.97479239
Tom Clancy Support and Defend - Mark Greaney	0.97337649
Commander-In-Chief - Mark Greaney	0.972952289
Locked On - Tom Clancy with Mark Greaney	0.972872574

…	…
Tom Clancy Under Fire - Grant Blackwood	0.954270016
MD09	0.949163736
MD06	0.93494537
MD05	0.933137011

TABLE 11: Cosine similarity scores for reference text “The Casual Vacancy” using full text comparison.

The Casual Vacancy - JK Rowling Ref text	**Stop word Cosine Similarity**
The Cuckoo's Calling - Robert Galbraith	0.993538902
The Silkworm - Robert Galbraith	0.985879318
Career of Evil - Robert Galbraith	0.98507601
Tom Clancy Full Force and Effect - Mark Greaney	0.978266791
Harry Potter 7 - JK Rowling	0.974386214
Harry Potter 4 - JK Rowling	0.973435664
Threat Vector - Tom Clancy with Mark Greaney	0.973255373
Without Remorse - Tom Clancy	0.97320388
Tom Clancy Support and Defend - Mark Greaney	0.972549814
Commander-In-Chief - Mark Greaney	0.972107585
…	…
MD07	0.925705929
MD09	0.918963822
MD05	0.902815078
MD01 ht2016 - Helen	0.900279023

TABLE 12: Cosine similarity scores for reference text “The Casual Vacancy” using stop word only comparison.

The Casual Vacancy attributes fairly poorly to the Harry Potter books. This attribution increases slightly when considering stop words only, but again, remains fairly weak, with only 4/10 and 5/10 attributing correctly for full text and stop words respectively. However, Robert Galbraith attributes extremely well to *The Casual Vacancy*. Once again, like many of the other experiments above, there was a small but not insignificant improvement to the authorship attribution when using stop words only.

The next part of this experiment examined the Tom Clancy books, and whether it was possible to differentiate between the novels written by Tom Clancy himself, and those written either in collaboration with another author, or written by another author entirely. We selected one of Mark Greaney's co-authored books, as well as one entirely authored by Clancy.

First, Tom Clancy books written entirely by Tom Clancy are shown in Table 13 and Table 14.

Debt of Honor - Tom Clancy Ref text	**Cosine Similarity**
The Cardinal of the Kremlin - Tom Clancy	0.996010335
Red Rabbit - Tom Clancy	0.995435781
The Hunt for Red October - Tom Clancy	0.994600684
The Teeth of the Tiger - Tom Clancy	0.994152975
Without Remorse - Tom Clancy	0.993179134
Patriot Games - Tom Clancy	0.992953165
Dead or Alive - Tom Clancy Grant Blackwood	0.992068612
Locked On - Tom Clancy with Mark Greaney	0.989949672
Threat Vector - Tom Clancy with Mark Greaney	0.98986492
Command Authority - Tom Clancy with Mark Greaney	0.989819969

TABLE 13: Cosine similarity scores for reference text "Debt Of Honor" using full text comparison.

Debt of Honor - Tom Clancy Ref text	**Stop word Cosine Similarity**
Executive Orders - Tom Clancy	0.999315266
The Bear and the Dragon - Tom Clancy	0.997922178
Rainbow Six - Tom Clancy	0.997850674

Clear and Present Danger - Tom Clancy	0.997705816
The Cardinal of the Kremlin - Tom Clancy	0.99635093
The Hunt for Red October - Tom Clancy	0.995970442
Red Rabbit - Tom Clancy	0.995523384
The Sum of All Fears - Tom Clancy	0.994599
Without Remorse - Tom Clancy	0.993651904
The Teeth of the Tiger - Tom Clancy	0.993053187

TABLE 14: Cosine similarity scores for reference text "Debt Of Honor" using stop word only comparison.

These results are quite encouraging. The top ten results for full text displays the strongest attribution to Tom Clancy books, and additionally, the top 6 works are for sole-authored Tom Clancy books. Moving to stop words only further increases the success of the authorship attribution, with the top ten most similar texts all being texts solely written by Tom Clancy. Indeed, the full results (available on request) shows of the 11 other texts authored by Tom Clancy, binary n-grams ranked all 11 as the most similar above all other texts.

While binary n-grams, especially with stop words only, attributed well for sole authorship for Tom Clancy, we also wanted to test how well they performed for joint authorship between Tom Clancy and a co-author. Mark Greaney was selected as the co-author to test, as, after Tom Clancy, he has written the most Tom Clancy universe books. Table 15 shows the results.

Command Authority - Tom Clancy with Mark Greaney Ref text	**Cosine Similarity**
Commander-In-Chief - Mark Greaney	0.99824592
Locked On - Tom Clancy with Mark Greaney	0.997327426
Threat Vector - Tom Clancy with Mark Greaney	0.996772952
Tom Clancy Full Force and Effect - Mark Greaney	0.996277013
Tom Clancy Support and Defend - Mark Greaney	0.993854313
Dead or Alive - Tom Clancy Grant Blackwood	0.990975916
The Cardinal of the Kremlin - Tom Clancy	0.990058336
Debt of Honor - Tom Clancy	0.989819969

Rainbow Six - Tom Clancy	0.989305428
The Bear and the Dragon - Tom Clancy	0.989261091

TABLE 15: Cosine similarity scores for reference text “Command Authority” using full text comparison.

Measuring this co-authorship is more challenging that the single author tests, because Tom Clancy only co-wrote three of the Tom Clancy universe books with Mark Greaney before he passed away. Further, the number of sole-authored Tom Clancy universe books by Mark Greaney is also less, at an additional three novels, with a 4th novel that was published after this experiment. So, whilst there may be fewer total texts correctly attributed in the top 10 most similar results, this is largely as a result of there simply being fewer texts available to test and all five of the other Mark Greaney texts are listed as the top 5 most similar results for the reference text. Further, although the most similar result was not a Tom Clancy and Mark Greaney collaboration text, the next two most similar results correctly identify the collaboration texts.

Table 16 provides a nearly identical picture.

Command Authority - Tom Clancy with Mark Greaney Ref text	**Stop word Cosine Similarity**
Commander-In-Chief - Mark Greaney	0.998608231
Locked On - Tom Clancy with Mark Greaney	0.998526096
Threat Vector - Tom Clancy with Mark Greaney	0.997882888
Tom Clancy Full Force and Effect - Mark Greaney	0.997778792
Tom Clancy Support and Defend - Mark Greaney	0.995607293
Rainbow Six - Tom Clancy	0.992578925
Dead or Alive - Tom Clancy Grant Blackwood	0.991485081
The Hunt for Red October - Tom Clancy	0.991269191
Executive Orders - Tom Clancy	0.99099649
Debt of Honor - Tom Clancy	0.990716033

TABLE 16: Cosine similarity scores for reference text “Command Authority” using stop word only comparison.

Unlike the Tom Clancy sole-authorship texts, for these texts, moving from a full text to stop word only analysis provides almost no benefit. Indeed, the top 5 texts remain in the same order, with an incorrectly-attributed sole-authored text by Mark Greaney text as the most similar, followed by the two correct co-author texts, which was followed by other Tom Clancy universe books, albeit in a different arrangement as the full text results - however, the re-ordering of these other Tom

Clancy universe texts cannot be considered either good, bad, or relevant, because, due to the small sample size of Mark Greaney texts, we have already run out of “correct” answers for attribution.

As with *The Cursed Child* texts, we examined the similarity scores of the most similar text, in this case, *Commander-In-Chief*. This was also a useful additional test because it was sole-authored by Mark Greaney.

The results were as in Table 17.

Commander-In-Chief - Mark Greaney Ref text	**Cosine Similarity**
Command Authority - Tom Clancy with Mark Greaney	0.99824592
Locked On - Tom Clancy with Mark Greaney	0.997165253
Tom Clancy Full Force and Effect - Mark Greaney	0.996911891
Threat Vector - Tom Clancy with Mark Greaney	0.996677112
Tom Clancy Support and Defend - Mark Greaney	0.994158991
The Cardinal of the Kremlin - Tom Clancy	0.990230646
Debt of Honor - Tom Clancy	0.98979044
Dead or Alive - Tom Clancy Grant Blackwood	0.989653092
Rainbow Six - Tom Clancy	0.989270374
The Bear and the Dragon - Tom Clancy	0.989119998

TABLE 17: Cosine similarity scores for reference text “Commander-In-Chief” using full text comparison.

Table 17 shows that the top two most similar texts were incorrectly identified as co-authored. This shows that in this case the most similar relationship between the co-authored texts and this text was reciprocal, unlike *The Cursed Child*, where the relationship was unidirectional. However, all the Mark Greaney books were correctly ranked above other Tom Clancy universe texts and well above other non-Tom Clancy universe texts.

Switching to a stop word analysis showed much the same, in Table 18.

Commander-In-Chief - Mark Greaney Ref text	**Stop word Cosine Similarity**
Command Authority - Tom Clancy with Mark Greaney	0.998608231
Locked On - Tom Clancy with Mark Greaney	0.998526468

Tom Clancy Full Force and Effect - Mark Greaney	0.99843992
Tom Clancy Support and Defend - Mark Greaney	0.996695401
Threat Vector - Tom Clancy with Mark Greaney	0.996684769
Rainbow Six - Tom Clancy	0.990692547
The Cardinal of the Kremlin - Tom Clancy	0.990299608
Executive Orders - Tom Clancy	0.989893864
The Hunt for Red October - Tom Clancy	0.989642072

TABLE 18: Cosine similarity scores for reference text "Commander-In-Chief" using stop word only comparison.

Again, the top 3 texts remained the same, whilst the remaining Mark Greaney texts experienced some minor re-arrangement, which was mirrored by the remaining sole-authored Tom Clancy texts.

5.3 Section Summary

This experiment was fairly freeform and investigated a number of interesting areas and drew several conclusions from the investigations.

First, context matters when dealing with authorship attribution and we demonstrated it using two related texts, *The Cursed Child*, compared to the other Harry Potter books. Despite both being about the same topic, we found that because *The Cursed Child* was written as a play, it had a number of different formatting and style choices that significantly reduced the similarities generated. An example of the formatting differences include character names always being fully capitalized (e.g. "HARRY") in the play, compared to almost never in the Harry Potter novelization. It seems likely that this topic sensitivity would extend to areas such as the intended aim of the text, for example, a casual blog post and an academic paper, even on the same topic, are likely to demonstrate a lower cosine similarity because the word choices used in these two contexts are very different. For example, a blog post may contain word abbreviations, a casual tone, and spelling mistakes, whereas an academic document is unlikely to. We also found that the use of stop words was insufficient for detecting and removing context sensitivity, but a future investigation into smarter stop word lists, as well as style words may be able to address this.

Another conclusion drawn based on this topic sensitivity, was that it is possible, when using a kNN-based classifier, for texts to only attribute in one direction. Using *The Cursed Child* example, of the 10 nearest neighbours to this text, 7 of them correctly attributed to other Harry Potter and J.K. Rowling works. However, when looking at *The Cursed Child* from a reference perspective of its top attributing text, the first Harry Potter book, *The Cursed Child* was the 36th nearest neighbour, out of a possible 41 neighbours – an extremely poor attribution score.

We also examined how well binary n-grams could differentiate between texts within the same fictional universe but written by different authors, as well as being co-authored. To test this we used texts in the Tom Clancy universe written by Tom Clancy himself, by Mark Greaney, as well as texts written as a collaboration between Tom Clancy and Mark Greaney. The results showed that the authorship attribution techniques were good at identifying same-universe-different-author

texts, and demonstrated some ability to differentiate between sole-authored and co-authored texts, however with the small amount of co-authored texts available in our corpus, the results were inconclusive, and need to be tested further using a much bigger co-author corpus.

6. DISCUSSION

6.1 What The Experiments Have Shown

The purpose of this paper was to explore binary n-grams in a sequence of experiments to test their performance. The sequence was intended to firstly establish whether binary n-grams were competitive with alternative methods such as character n-grams and bag of words. While binary n-grams are conjectured to have useful characteristics such as maximising frequencies and language- and text encoding independence, they had not been compared to the most similar methods, but had only been assessed for feasibility as an authorship attribution method. The sequence then explored the effect of various word types being included or excluded, particularly function and stop words, and how different types of text and combinations of authors affected authorship attribution.

In section 2, the experiment tested binary n-grams against two other methods of authorship attribution, character n-grams and bag of words, on the Federalist Papers, a well-known authorship attribution problem dating back to 1787. The results found that binary n-grams performed at least as well as bag of words, and slightly better than character n-grams. Further, the results for character n-grams were in line with previous research [18].

In section 3, the experiment examined the effect of various n-gram sizes on authorship attribution. The results found that the worst performance of 50% correct attribution occurred with the smallest size n-grams, and that performance increased as the size of the n-gram increased, until a plateau was reached at 80% correct attribution starting at n = 15, which continued until = 18. There was a small spike of 83.33% attribution at n = 14, but it was uncertain whether this was an outlier or a significant finding.

Previous experiments agreed with the literature that content words seemed to have a significant influence on attribution, in that texts with similar topics by different authors were more likely to be attributed together, and that texts with dissimilar topics were less likely to be attributed together, even if they shared the same author. The section 4 experiment removed the content words from the texts before the binary n-gram analysis was performed. The results found that a modest increase of 3.33% in overall text attribution was achieved, and that the plateau of attribution, after which no further increase was observed, commenced at a lower n-gram size of n = 11. Examining these results, as well as drawing inferences from the texts themselves, as well as the list of stop words used, it seems that a list of stop words on its own was insufficient at removing topic sensitivity, and further investigation into “style” words was indicated for future work.

The section 5 experiment calculated a large number of cosine distances between texts, the majority of which were written by J.K. Rowling, or set in the Tom Clancy universe and written by either Tom Clancy himself, Mark Greaney, or co-authored by both Tom Clancy and Mark Greaney. This experiment yielded many interesting observations. The first and most important finding is that context seems to be very important in performing authorship attribution. Also, while excluding content words can help alleviate issues of topic sensitivity, it was not the only consideration. Results for the Harry Potter play, The Cursed Child attributed very poorly to topic-similar texts (the other Harry Potter books) because of contextual differences inherent to being in the play format, for example, having character names typed as all capitals, or having: “ACT ONE, SCENE ONE” style annotations throughout the text. This clearly demonstrated that cross-genre authorship attribution requires more research.

Finally, we examined whether binary n-grams can be used to detect and differentiate between sole-authored work and co-authored work when working with texts on the same topic. While the results did show some promise, because of the limited number of co-authored texts available to

use in the corpus, the results were inconclusive and further investigation is required to produce a more conclusive result.

6.2 Observations Arising

The experiments looked at whether a combination of binary n-gram analysis with content word removal can improve authorship attribution and found some scope for optimism. However, there are many remaining problems of authorship attribution.

There may be effective attribution methods for specific domains, specific languages, or even source types. However, optimising for one language may not contribute to others, especially if the alphabets are different. Also, such optimisation might give misleading results in documents containing text in multiple languages.

Two approaches could be made to work in such as case. The first is to make a special method for each language (or even each domain, cross-domain or single-domain). This could be adapted so that each specialist method could be used for each tract in a given language within any multi-language text (although identifying the language at a given point can be challenging). That said, treating languages separately means reducing the available text for each language which can be a problem for small texts.

The alternative is to use a method which operates independently of the language, its constructs, its alphabets and even whether there is more than one language present. The binary n-gram method does this, and the character n-gram method can also do so, although perhaps less efficiently for multiple alphabets, by treating them all together, ignoring the language. It also does this more efficiently than the character n-gram method in the above experiments, although the comparison has not yet been done across languages or for the much smaller text files that make up social media content.

A problem of interest is authorship attribution of social media content, contributed by unknown authors. Some authors hide behind pseudonyms to conceal their political or commercial motivations, masquerading as 'real people'. Some are not human. Being able to attribute those authors could expose astroturfers (using many accounts to create a false 'grass roots' movement) or sock puppets operating on behalf of agencies.

However, a major challenge of social media content is that it often comprises very small quantities of text, and often the language is less compliant with more traditional writing rules. Tweets exemplify this. These features make it challenging to use many of the otherwise highly effective author attribution methods. For example, if the text is written in informal language that does not observe spelling or grammatical rules, methods that rely on such features will fail. Methods that make well-informed assumptions about features such as prefixes may not detect enough of those features to operate adequately if the text is formed differently - every specialisation introduces a potential point of weakness. Even if they do work, the brevity of the texts may render them inadequate just because there is not enough data.

The section 4 experimentation considered a method that reduces text as a side effect. Leaving out content words may reduce the misattribution that occurs due to the topic, rather than the author, influencing the attribution. This worked well enough in the large-file attributions in the experiments.

However, excluding text of any sort at all comes with the concomitant reduction of text for analysis. This means a reduction in analysable features, which may totally compromise the analysis if trying to attribute small texts, such as social media postings, especially tweets. So when author-attributing social media content, any method which reduces the quantity of text available is a problem, because there is already a scarcity of text for analysis.

There are two approaches to dealing with shortage of text for analysis. The first is to just not exclude anything. However, this leaves the analysis potentially weakened, if not actually compromised, by leaving in the full text. Certainly the experiment in section 4 indicates that attribution would be less accurate if leaving in content words.

The second approach is to make maximum use of what little text may be available. This could be implemented by seeking out language-specific features, although it could come at the cost of being able to analyse multi-language texts. It could also be done by minimising the units over which analysis is done to make it possible to extract more frequencies from less text, which is what binary n-gram analysis offers over character n-gram analysis. In section 2 the performance of binary 14-grams was found to be better than that of character 4-grams. Dividing a short text of, say, 280 characters into 14-bit segments yields 1946 values, but dividing it into 4-character segments yields only 276 values. While there is not actually more text data present than with character n-grams, more frequencies are extracted from it and, as section 2 indicated, appeared to give better performance. A future experiment will be to compare binary versus character n-grams over a set of short texts, to establish whether the better performance is sustained in that context.

6.3 Social Media Content and Authorship Attribution

Challenges of authorship attribution become more apparent when attempting to analyse new forms of text, in particular, social media. Social media is now too large a facet of authorship attribution to ignore, with sites like Facebook and Twitter being respectively the third and twelfth most popular sites globally [28] [29]. In addition to the social networking, social media now performs other functions, like news updates that can spread at breakneck speeds. Platforms like Twitter were used extensively by journalists, police, and individuals during emergencies like the Boston Marathon Bombings where news about the attack, safety information, and images of the alleged attackers were shared over Twitter [30].

It is also hard to overstate the importance of applying authorship attribution to social media platforms, either; for example, in 2015 the Guardian newspaper wrote an expose on workers interviewed from a “Russian troll” shill posting ring that comprised of “hundreds of bloggers” being paid, unofficially, by their government to work around the clock posting pro-Kremlin and anti-Western messages to community forums and online blogs [31]. Although eventually discovered by other means, authorship attribution is also extremely relevant to this example – if a large number of pro-Kremlin posts across several social media accounts could be attributed back to a single author or group of authors it would provide the ability to discriminate between “real” messages and blogs, and “fake” messages and blogs paid for by a third party with an agenda.

Despite the value of performing authorship attribution on social media platforms, they present unique and significant challenges for any current method of authorship attribution. The need for new methods for authorship attribution is clear, with traditional techniques' effectiveness on platforms like Twitter, where each Tweet must be delivered in 280 characters or fewer, failing to reach 50% according to a recent review into authorship attribution on social media, with recommendations noting “a significant need in forensics for new authorship attribution algorithms” that are effective on such platforms [11].

As noted above, many of the algorithmic improvements for authorship attribution rely on ignoring or removing parts of the text. In the context of very small texts being attributed, this can reduce the amount of text beyond accurate attribution.

While the binary n-gram approach may offer an effective general approach to authorship attribution of small texts, it too will have its lower bounds beyond which it cannot operate. Perhaps the only sensible solution is an ensemble approach that applies a wide range of complementary authorship attribution methods and deduces an outcome based on many different characteristics. The binary n-gram approach is one such complementary approach, offering

unique strengths in terms of language and alphabet independence and could contribute one further method to any ensemble of authorship attribution tools.

7. CONCLUSIONS AND FUTURE WORK

This paper considered the combination of the binary n-gram method for authorship attribution with the technique of reducing the topic-sensitivity of authorship attribution by excluding all content words. The structured sequence of experiments firstly established that binary n-gram analysis was as good as the established character n-gram analysis, which is claimed in the literature to be among the best of authorship attribution methods. It then determined what the appropriate size of n-gram for binary n-gram analysis was, with 14-grams best for full text and 11-grams best for text excluding content words. Alternative authorship attribution scenarios were then explored, such as including cross-domain and cross-genre works, with the final experiment including a scenario where multiple authors contribute to works in a single-domain.

The experiments indicated that the combination of binary n-gram analysis with removal of content words was reasonably effective. However when considering the growing need for authorship attribution to be carried out over social media content, any methods that reduce the text available from already-small files may not be effective. It remains to be determined how well these and other methods operate over social media content and what the absolute limits of authorship attribution may be.

8. REFERENCES

[1] Judges 5:5-6. Holy Bible. Authorised King James Version.

[2] T. Merriam. "Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe". *Literary and Linguistic Computing*, vol. 9 (1), pp. 1-6. 1994.

[3] R. Matthews. "Neural Computation in Stylometry I: An Application to the Works of Shakespeare and Fletcher". *Literary and Linguistic Computing*, vol. 8 (4), pp. 203-210. 1993.

[4] D. Lowe and R. Matthews. "Shakespeare vs. Fletcher: A stylometric analysis by radial basis functions". *Computers and the Humanities*, vol. 29 (6), pp. 449-461. 1995.

[5] A. Hamilton, J. Madison, J. Jay and J. Rakove J. "The Federalist". Bedford/St. Martin's, Boston. 2003

[6] E. Stamatatos. "A survey of modern authorship attribution methods". *Journal of the American Society for Information Science and Technologies*, vol. 60 (3), pp. 538-556. 2009.

[7] P. Juola. "Authorship Attribution". *Foundations and Trends in Information Retrieval*, vol. 1, (3), pp. 233-334. 2006.

[8] H. Fouche Gaines. H. *Cryptanalysis*. Dover, New York. 1956.

[9] M. Kestemont. "Function Words in Authorship Attribution: From Black Magic to Theory?". *Proc. 3rd workshop on Computational Linguistics for Literature*, pp. 59-66, Gothenburg, Sweden, ACL, https://www.aclweb.org/anthology/W14-0908 2014,

[10] E. Stamatatos. "On the robustness of authorship attribution based on character n-gram features". (Symposium: Authorship Attribution Workshop). *Journal of Law and Policy*, vol. 21, pp. 421-439. 2013.

[11] A. Rocha, W. Scheirer, C. Forstall, T. Cavalcante, Theophilo, B. Shen, A. Carvalho and E. Stamatatos. "Authorship Attribution for Social Media Forensics". *IEEE Transactions on Information Forensics and Security*, vol. 12 (1), pp. 5-33. 2017.

[12] J. Peng, S. Detchon, K-KR. Choo and H. Ashman. "Astroturfing Detection in Social Media: A Binary N-gram Based Approach". *Concurrency and Computation: Practice and Experience*, doi: 10.1002/cpe.4013. 2016.

[13] J. Peng. "Authorship Attribution with Binary N-gram Analysis for Detecting Astroturfing in Social Media". PhD thesis, University of South Australia, Australia. 2017.

[14] HDJ. Coupe. "Non-Symbolic Fragmentation Cryptographic Algorithms". PhD thesis, University of Nottingham, UK. 2005.

[15] U. Sapkota, S. Bethard, M. Montes-y-Gómez and T. Solorio. "Not all character n-grams are created equal: A study in authorship attribution". *Proc. Annual Conf. North Amer. Chapter ACL Human Lang. Technologies*. https://www.aclweb.org/anthology/N15-1010, pp. 93-102. 2015.

[16] K. Sundararajan and D. Woodard. "What constitutes 'style' in authorship attribution?". *Proc. 27th Int. Conf. on Computational Linguistics*. Assoc. Computational Linguistics. pp. 2814–2822, https://www.aclweb.org/anthology/C18-1238. 2018.

[17] J. Peng, K-KR. Choo and Ashman H. "Bit-level n-gram based forensic authorship analysis on social media: Identifying individuals from linguistic profiles". *Journal of Networked and Computer Applications*, vol. 70, pp. 171-182. 2016.

[18] B. Kjell, W. Woods and O. Frieder. "Discrimination of authorship using visualization". *Information Processing and Management*, vol. 30 (1), pp. 141-150.

[19] US Congress. "The Federalist Papers". Congress.gov Resources. (accessed 2019/09/10), 2017. https://www.congress.gov/resources/display/content/The+Federalist+Papers.

[20] V. Kešelj, F. Peng, N. Cercone and C. Thomas. "N-gram-based author profiles for authorship attribution". *Proc. of the Pacific association for computational linguistics*, Vol. 3, pp/ 255-264). 2003.

[21] R. Galbraith. "About Robert Galbraith". 2019/07/25, http://robert-galbraith.com/about/. 2017. (accessed 2019/09/10).

[22] D. Doyle. "Stopwords" (English) (accessed 2019/09/10), http://www.ranks.nl/stopwords. 2017.

[23] L. Milos. "Playing the Pronoun Game: Are All of The Hobbit's Dwarves Male?". http://middleearthnews.com/2018/01/09/playing-the-pronoun-game-are-all-of-the-hobbits-dwarves-male/ (accessed 2019/09/10). 2018.

[24] B. Blatt. *Nabokov's favourite word is Mauve*. Simon and Schuster. 2017.

[25] J. Rowling, J. Tiffany J and J. Thorne. *Harry Potter and the cursed child*. Little & Brown, London. 2016.

[26] J. Rowling. *Harry Potter and the Half-Blood Prince*. Pottermore, England. 2012.

[27] T. Clancy. *Locked On*, by Tom Clancy with Mark Greaney. (accessed 2019/09/10), https://tomclancy.com/product/locked-on/. 2017.

[28] Alexa. "Facebook.com Traffic, Demographics and Competitors". (accessed 2019/09/10), 2019. https://www.alexa.com/siteinfo/facebook.com.

[29] Alexa. "Twitter.com Traffic, Demographics and Competitors". (accessed 2019/09/10), 2019. https://www.alexa.com/siteinfo/twitter.com.

[30] S. Rogers. "The Boston Bombing: How journalists used Twitter to tell the story". (accessed 2019/09/10), https://blog.twitter.com/official/en_us/a/2013/the-boston-bombing-how-journalists-used-twitter-to-tell-the-story.html. 2017.

[31] S. Walker. "Salutin' Putin: inside a Russian troll house". (accessed 2019/09/10), https://www.theguardian.com/world/2015/apr/02/putin-kremlin-inside-russian-troll-house. 2017.