CLOSE AND OPEN TASK AUTHORSHIP ATTRIBUTION: A COMPUTATIONAL AUTHORSHIP ANALYSIS

Nur Inda Jazilah
Faculty of Humanities, Vrije Universiteit,
Amsterdam, Netherland
nurindajazilah@student.vu.nl

Abstract

Authorship analysis is an area lies within forensic linguistics where the main task is to investigate the characteristics of a text in terms of its authorship. Specifically, authorship attribution examines the possibility of an author for writing the text by analyzing the author's other works. This experimental research addresses two problems: which author writes which text (using a closed task authorship attribution) and who writes each text (using an open task of authorship attribution). In doing so, this research uses R to do statistical computing and employs both stylo and classify functions. Based on carried out experiments with a fixed 1-gram variable, it is concluded that SVM algorithm may be best used in doing closed-task authorship attribution for its 100% consistency, whereas for the open task k-NN algorithm may be best used if it reaches 94% consistency. In addition, stylo function may perform better than classify function since style function provides closer to the actual answer results. Scientifically, this research provides a framework of how to do authorship analysis computationally and practically it is projected as a tool to detect plagiarism.

Keywords: Authorship Analysis; Computational Approach; Forensic linguistics; Classify Function; Stylo Function

INTRODUCTION

Juola (2008) defines authorship attribution as 'the science of inferring characteristics of the authors from the characteristics of documents written by that author'. Luyckx (2010) adds that authorship attribution aims to identify the author of the unknown text(s) by analyzing available text(s) written by a number of candidate authors. It is noticeable that research about authorship analysis has been under-investigated. Authorship analysis employs stylistics to analyze a writing style of an author which later can be used as a base to compare with questioned texts. This methodology may be helpful in analyzing plagiarism to see whether a different author writes text or not. As plagiarism cases have been often found in Indonesia – even several academicians such as lecturers and chancellors from a reputable university have been found committed plagiarism, this study may contribute to plagiarism detection in a hope to decrease the number of plagiarisms.

Several studies have been done in the area of authorship analysis with different object of studies – ranging from computer programs (Gray, MacDonell, and Sallis, 1997), social media (Peng, Choo, and Ashman, 2016), text messages (Grant, 2010), email (de Vel et al., 2001), to textual communications (Iqbal et al., 2013). Two other studies on how to do authorship analysis also have been done by Grant (2007) discussing how to quantify evidence in authorship analysis and Grant (2008) concerning on how to approach questions in authorship analysis. Furthermore, Zheng et al. (2003) propose a practical value of authorship analysis in cybercrime investigation.

This research is designed to solve the problem of authorship verification in which a text classification is carried out by assigning the tool to determine whether or not the text(s) is written by author X for example. Luyckx (2010) explains further that in this authorship attribution, an open candidate set is given as well as to the Stylo tool. This set excludes negative examples such as text(s) that has not been written by the author X. The experiments in this research is an attempt to address problems that often emerge in the field of forensic linguistics, that is about how to determine whether a text was written by one of the persons listed in the police's list of suspected persons for instance. This research will address the following research questions: (1) performing a closed task authorship attribution, which author does write which text? And (2) performing an open task authorship attribution, who writes each text provided in the test set?

Addressing the aforementioned research questions, this research aims at providing both theoretical and practical values. Scientifically, it is expected that this paper will contribute in detecting plagiarism within the area of authorship analysis. Besides, this research will make provision for a framework on how to carry out authorship analysis computationally, specifically using R – one of the programming languages by utilizing Stylo and Classify functions. In regard to practical values, the framework can be used to analyze the possible author form a list of suspected authors. In other words, these two functions can be used as a tool to detect plagiarism by attributing a text to which author.

This paper will sequentially discuss the following sections in order: a theoretical framework on which this study is based, the methodology used to address the abovementioned research questions including the data and the tool used in this study, the results of the experiments, and conclusions.

THEORETICAL FRAMEWORK

Authorship Analysis

Authorship analysis is measuring an author's style of writing which can be done using stylometry. Verhoeven (2015) defines stylometry as 'the quantitative study of stylistic characteristics of a text.' Authorship analysis can be done either qualitatively or quantitatively. Qualitative analysis may be performed by looking at linguistic features portrayed in the text(s) utilizing a stylistic method which can go either way starting from the known texts to the unknown ones or the other way around. McMenamin (2002) defines linguistic stylistics as 'the scientific interpretation of style-markers as observed, described, and analyzed in the language of groups and individuals. It aims at seeking information regarding a group's membership and patterns that distinguish a member from the group via style markers. Although McMenamin (2002) has pointed out several style markers through his analysis of 80 authorship analysis cases, it is evident that these markers may be not up to date considering today's modes of communication such as text messages and social media (Oliveira, van der Voet, and Jazilah, 2018). It is primarily because those two modes of communication are typed rather than handwritten. However, Smith, Spencer, and Grant (2012, as cited in MacLeod and Grant, 2012) adds several style markers to McMenamin's list – among the markers are (1) strikeovers and cross-outs, (2) mistakes, errors, and typos, (3) structure of information, (4) smileys or emoticons, hashtags, and mentionings, and (5) country-specific.

Apart from qualitative analysis, authorship analysis also can be done quantitatively. McMenamin (2002) argues that statistical analysis may be necessary to measure how many variations and how often they are used. Several tests that may be carried out are frequency distribution, means (such as standard error of difference, t-test, and analysis of variance), percentages (i.e., proportion test), frequencies (i.e., chi-square), variable independence using coefficient correlation, and joint probability using a frequency estimate. It is expected that through this quantitative analysis, forensic linguists may help the juries or judges to provide a more informed decision (McMenamin, 2002) as well as make forensic science more scientific (Solan, 2010). What may be challenging in doing quantitative analysis relates to the richness of the data – since the data in a criminal case are likely small data, therefore, it may be difficult to quantify. Moreover, Grant (2013) only uses features that appear 'at least twice as many messages as the other'.

In addition to qualitative and quantitative analysis, authorship analysis can be done conventionally and computationally. In conventional authorship analysis, a forensic linguist performs manual analysis by comparing features in both known and questioned texts. This analysis may start from the known texts to the unknown ones or vice versa. On the other hand, computational authorship attribution analyzes the text(s) with the assistance of a tool in a computer. When the text(s) is considerably large, a computational method may be best used. Olsson (2009) exemplifies a conventional qualitative authorship analysis on text messages. He pointed out who the murderer is after analyzing several text messages sent to the phone of victim's husband, the perpetrator's letter, and a police interview of the suspect. Evidence that drives him to say so is the use of a period instead of a comma and the use of word sort out in quite rare contexts.

To illustrate computational authorship analysis, Hughes (2013) narrates that Juola and Millican run an analysis to prove if JK Rowling writes The Cuckoo's Calling. Juola ran an analysis through a computer program Java Graphical Authorship Attribution Program (JGAAP) for approximately an hour and a half to see word pairings and character n-grams of the novel. He also points out that the word length of the novel is 'very characteristically Rowling' which serves as the strongest evidence that Rowling writes the novel. Meanwhile, Milican ran a parallel investigation on Rowling using his program, Signature. The program involves a statistical method, principal component analysis, covering six features: word length, sentence length, paragraph length, letter frequency, punctuation frequency, and word usage. Through word usage, Millican found that 'on the graph, it's absolutely clear that Cuckoo's Calling is lining up with Harry Potter' (Hughes, 2013).

METHOD

To be able to address the problem, this research requires data that will be analyzed as well as a research methodology. Two following issues need to consider in doing authorship analysis – scalability and choice of qualitative or quantitative analysis. Luyckx (2010) discusses scalability issues extensively in her dissertation. She remarks that scalability issues may be viewed from different aspects ranging from feature selection, techniques, and validation choice, the author set size, training data size, and approach. In relation to the choice of qualitative or quantitative analysis, Solan (2013) asserts that courts tend to prefer quantitative analysis since it has a lower risk of cognitive prejudice. However, Grant (2010) argues that a holistic quantitative analysis maybe not possible in forensic linguistics and suggests linguists work with integrity in all cases. Thus, a combination of qualitative and quantitative analyses may be useful in dealing with cases enabling for both – using quantitative techniques when the data enable it and adds to the analysis (Oliveira, van der Voet, and Jazilah, 2018). This section will elaborate on the data and method used to solve the problem in this research.

The data in this research is relatively small and in the form of English texts which is accessible through the web of PAN competition[1]. The texts are narrative texts where the length of the texts is various. Pre-processing analysis has not been done to the data so that it still contains some punctuations such as quotation and exclamation marks, apostrophe, etc. The data are divided into three parts: (1) training corpus, (2) test corpus closed class, and (3) test corpus, open class. The training corpus is a collection of sixteen texts from eight authors – two texts represent each author's style, so there are sixteen texts respectively. The test corpus closed class consists of eight texts from unknown authors, while the test corpus open class comprises of seventeen texts from unknown authors as well. Both these corpora are later used as secondary sets to test them. Subsequent to running classify in the training set, a style of each author is extracted. This writing style is used later as such guide to test both the secondary sets of the closed and open class task to be able to determine who the writer of text is.

[1] The data are accessible at https://pan.webis.de/clef12/pan12-web/author-identification.html

This research lies within the area of Natural Language Processing that aims to detect a pattern of texts[2]. What is meant by a pattern here is noticeable features extracted from the texts that later are used to classify the test sets. The methodology used in this research is text classification by machine learning. The given data, specifically the training corpus, is used as the dataset to extract features of given texts. This process of formulating a writer's style employs computational stylometry with the assistance of machine learning, i.e., attributing or profiling an author based on the measurement of the style that they use in writing (Eder, Rybicki, and Kestemont, 2016). The machine then will use the information of the features of each author in the training set to both test corpora to determine who the author of a text is. Two tasks performed in this research are attributing authorship for (1) closed and (2) open class. In closed class, the task is to define who writes what, meaning that each author in the training set writes a text in the test set. On the other hand, in open class, the task is different from the previous one since there may be a chance in which an author does not write any of the texts in the test set.

To be able to perform the task, this research uses Stylo tool which can be downloaded from Github, an online platform of software development[3]. Stylo is 'a flexible R package for the high-level analysis of writing style in stylometry' (Eder, Rybicki, and Kestemont, 2016, p. 107). This task is performed by merely opening R and then calling the library Stylo. Once the library can be successfully accessed, the next step is to set the directory where the datasets are, and the results will be stored. After setting the directory, the next step is to perform the classify() function that is used to classify which author writes which text.

To obtain the results and address the problem in this research, both closed and open class are tested in a quite similar way. The closed class is tested using five algorithms provided by the tool: Delta, k-NN, SVM, NaiveBayes, and NSC. The dataset is tested four times in every algorithm: two times using n-grams with the n = 1 and n = 3. These four-time tests are carried out to examine whether differences emerge when different features are chosen. As an illustration, four tests are performed using Delta algorithm. The first two tests use n = 1 with one test does not select the option of the preserved case, and the other one selects it. It is done to find out whether Delta algorithm is case sensitive or not in assigning text to an author. The other two tests use n = 3 and apply the same option of preserving case or not, as in the first two tests. These four tests using a different number of n are performed to detect whether the number of n affects the result. These four-time tests are performed in all five algorithms. Moreover, these parameters are also applied in the open class.

In addition to that, both classes are tested using similar parameters which are further explained in the followings:

[2] The details about what Natural Language Processing is can be found at http://www.nltk.org/book/ch06.html

[3] The tool can be downloaded from https://github.com/computationalstylistics/stylo

a. in input and language tab, plain text is chosen for the input since the texts are a kind of prose in a .txt type
b. for the language, English(ALL) is chosen to make sure that the contractions in the texts are treated as one word
c. UTF-8 is checked for the texts to be processed for those who use Mac
d. in the feature tab, words are chosen with the MFW minimum is 100, maximum 100, and increment is 100, and the culling minimum is 0, maximum is 0 and increment is 20
e. in the statistics tab, all algorithms are tested one by one with the option general as ALL culling, and these experiments use Classic Delta as it is considered a good choice for English texts.

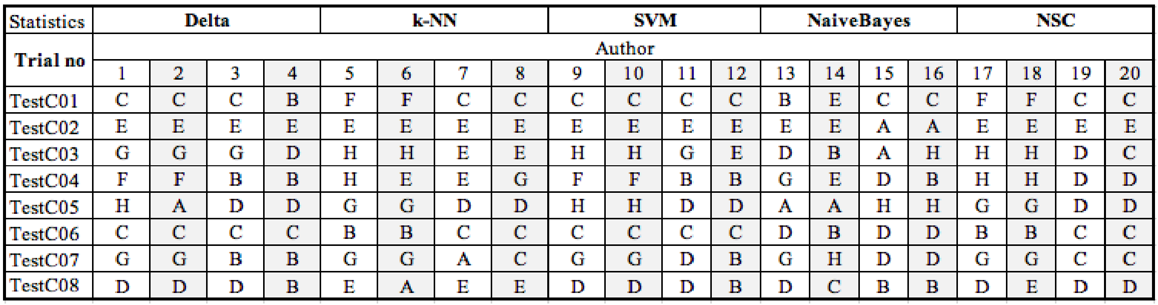
FINDINGS AND DISCUSSION

This section will elaborate on the result of the experiments that are divided based on the tasks. The first subsection will discuss the result of the performed task in attributing authorship in a closed class. The following subsection will address the open class one.

Authorship Attribution Closed Class

After running all twenty experiments, the results of attributing an author to a text in the closed class test set in all algorithms are shown in the following table.

Table 1: The result of experiments done in closed-class dataset

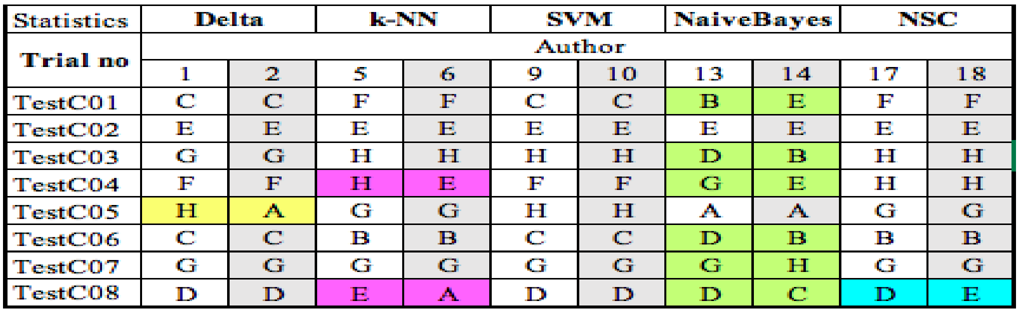


Statistics	Delta				k-NN				SVM				NaiveBayes				NSC			
Trial no	Author																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
TestC01	C	C	C	B	F	F	C	C	C	C	C	C	B	E	C	C	F	F	C	C
TestC02	E	E	E	E	E	E	E	E	E	E	E	E	E	E	A	A	E	E	E	E
TestC03	G	G	G	D	H	H	E	E	H	H	G	E	D	B	A	H	H	H	D	C
TestC04	F	F	B	B	H	E	E	G	F	F	B	B	G	E	D	B	H	H	D	D
TestC05	H	A	D	D	G	G	D	D	H	H	D	D	A	A	H	H	G	G	D	D
TestC06	C	C	C	C	B	B	C	C	C	C	C	C	D	B	D	D	B	B	C	C
TestC07	G	G	B	B	G	G	A	C	G	G	D	B	G	H	D	D	G	G	C	C
TestC08	D	D	D	B	E	A	E	E	D	D	D	B	D	C	B	B	D	E	D	D

The table illustrates the results of four-time experiments done on the test set using different five algorithms. Delta algorithm, for example, has the trial number from 1 to 4 with the following details: 1) the test with n = 1 and the option of preserved case is not selected, 2) the test with n = 1 and the option of preserved case is selected, 3) test using n = 3 and the option of preserved case is not selected, and 4) test using n = 3 and the option of preserved case is not selected. The trials with an odd number are the ones that do not select the option of the preserved case (in white color), and those with the even number are the ones that select the options of the preserved case (in grey color). The following trial number continues the trial number to ease documenting the results.

The following table is derived from the previous table comparing the results of trials in five different algorithms using the same number of n, i.e., n = 1.

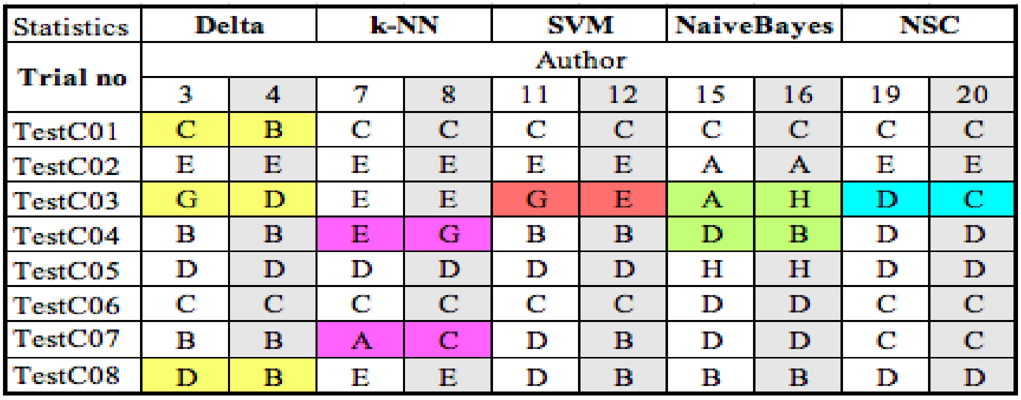
Table 2: The result of experiments done with n-grams = 1



Statistics	Delta		k-NN		SVM		NaiveBayes		NSC	
Trial no	Author									
	1	2	5	6	9	10	13	14	17	18
TestC01	C	C	F	F	C	C	B	E	F	F
TestC02	E	E	E	E	E	E	E	E	E	E
TestC03	G	G	H	H	H	H	D	B	H	H
TestC04	F	F	H	E	F	F	G	E	H	H
TestC05	H	A	G	G	H	H	A	A	G	G
TestC06	C	C	B	B	C	C	D	B	B	B
TestC07	G	G	G	G	G	G	G	H	G	G
TestC08	D	D	E	A	D	D	D	C	D	E

As can be seen in the above table, several cells are painted in different colors to show different results within 1-grams experiments between choosing the option of the preserved case and not. For Delta algorithm, there is one inconsistent result between both tests as well as in the NSC algorithm. The k-NN algorithm shows two different results between both tests. The most striking difference can be seen in the NaiveBayes algorithm, the results differ in six out of eight tests. The trials using the SVM algorithm shows a consistent result. A result of an algorithm is considered consistent if two tests using the same number of n provide an exactly similar result regardless of the option of the preserved case. On the other hand, it is vice versa –the result is considered as inconsistent if the two tests with the same number of n in a similar algorithm, the tool assigns text to the different author.

Table 3: The result of experiments done with n-grams = 3



Statistics	Delta		k-NN		SVM		NaiveBayes		NSC	
Trial no	Author									
	3	4	7	8	11	12	15	16	19	20
TestC01	C	B	C	C	C	C	C	C	C	C
TestC02	E	E	E	E	E	E	A	A	E	E
TestC03	G	D	E	E	G	E	A	H	D	C
TestC04	B	B	E	G	B	B	D	B	D	D
TestC05	D	D	D	D	D	D	H	H	D	D
TestC06	C	C	C	C	C	C	D	D	C	C
TestC07	B	B	A	C	D	B	D	D	C	C
TestC08	D	B	E	E	D	B	B	B	D	D

The results between those that are tested using 1-grams and 3-grams are different. When 3-grams is used, Delta algorithm is more inconsistent looking at the increased number of inconsistency result, from one to three. The k-NN algorithm shows consistency in terms of its correct number in attributing the authors: two out of eight are inconsistent. It is interesting to notice that the 3-grams NaiveBayes algorithm results in the opposite of the 1-grams showing only two inconsistencies. SVM algorithm – which previously results on a consistent attribution across all eight tests – and NSC algorithm have one inconsistent result.

Based on the mentioned results, it can be concluded that in 1-grams experiments SVM performs best showing 100% consistent result while NaiveBayes performs worse showing only 25% consistent result. On the other hand, in 3-grams tests both SVM and NSC performs better with 87,5% consistent result whereas Delta performs not entirely well with 62,5% consistent result. Thus, the result shown by SVM in 1-grams test is used as the final result. It is primarily because although SVM and NSC result on the same accuracy, SVM is chosen because it performs better than the NSC regarding consistency in 1-grams words. Thus, the final result will be based on the result of SVM with 1-grams test.

However, it is interesting to notice that author C and H appear to write two texts of the test set. In addition, the attributions based on these experiments differ in two test texts –cells in green color: testC03 and testC06 (see table 4). The table shows that both testC03 and test testC05 are written by the same author H, and testC01 and testC06 are also written by the same author C. This may be because the style of those pairs of text is close to either author H and C. To make sure which one is which, a conventional authorship analysis may need to be done to compare the two texts and define the author.

Table 4: the comparison between the experiments result in this study with the actual answers

File Name	Results of Experiments	Actual Answers	File Name	Results of Experiments	Actual Answers
testC01	C	C	testC05	H	H
testC02	E	E	testC06	C	B
testC03	H	A	testC07	G	G
testC04	F	F	testC08	D	D

As a comparison, an additional experiment using stylo function is done. This task is based on cluster analysis, and the result can be seen in figure 1. It can be seen that some of the classifications are different. Stylo function assigns text to an author as fifty percent same as the experiments –it assigns authorship differently in testC03, testC04, testC05 and testC06. It also assigns three different attributions compared to the provided answers: testC03, testC04, and testC05. In cluster analysis, it is noticeable that testC05 is clustered with one of A's writings but in the same level with H's writings which may lead to the attribution of testC05 to H. TestC03 is paired with textC07 and one of G's writings in one branch which may be because the testC03 has features that close to the training text G_1 and test set testC03. The other difference is the attribution of testC04, that both of the experiments and actual answers match the text with the author F. However, the cluster analysis groups the text with one of the B's writings.

DOI: http://dx.doi.org/10.18860/prdg.v2i1.6704

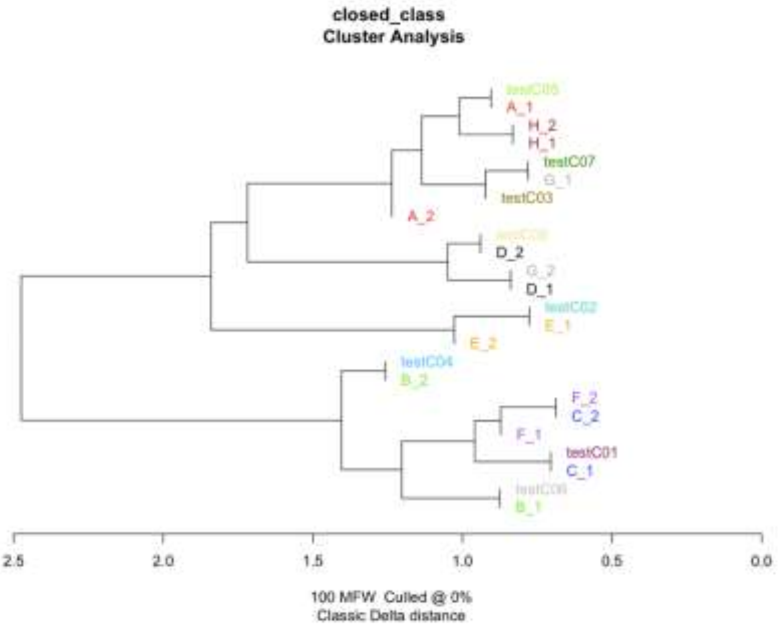


Figure 1: the result of stylo function in closed class

Authorship Attribution Open Class

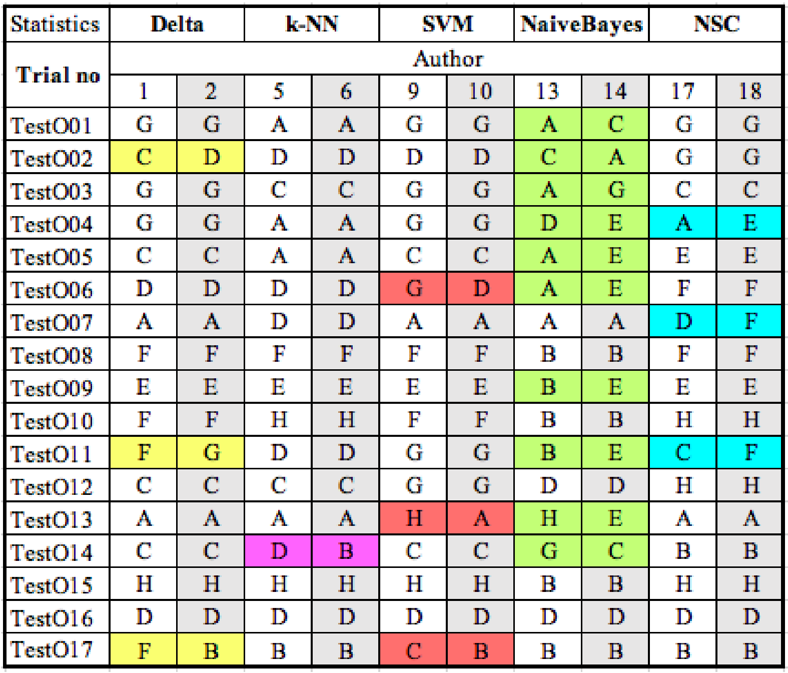
Twenty experiments are carried out in open class test set to determine which text is written by whom as illustrated in table 5. The experiments' results are presented similarly to how the closed class results are shown in the previous subsection.

Table 5: The result of experiments done in open class dataset

Statistics	Delta				k-NN				SVM				NaiveBayes				NSC			
Trial no	Author																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
TestO01	G	G	B	B	A	A	D	B	G	G	D	D	A	C	B	B	G	G	B	B
TestO02	C	D	G	G	D	D	E	G	D	D	B	G	C	A	B	B	G	G	D	D
TestO03	G	G	G	G	C	C	G	G	G	G	D	D	A	G	B	B	C	C	D	D
TestO04	G	G	A	E	A	A	A	A	G	G	E	E	D	E	A	A	A	E	C	C
TestO05	C	C	B	B	A	A	E	E	C	C	E	E	A	E	C	C	E	E	C	B
TestO06	D	D	G	G	D	D	E	G	G	D	G	G	A	E	B	B	F	F	G	G
TestO07	A	A	G	G	D	D	G	G	A	A	G	D	A	A	B	B	D	F	D	D
TestO08	F	F	G	G	F	F	E	G	F	F	E	G	B	B	B	B	F	F	E	E
TestO09	E	E	E	B	E	E	E	E	E	E	E	E	B	E	C	E	E	E	E	E
TestO10	F	F	E	B	H	H	E	E	F	F	E	D	B	B	F	H	H	H	D	D
TestO11	F	G	B	F	D	D	E	F	G	G	B	F	B	E	D	F	C	F	D	F
TestO12	C	C	G	G	C	C	E	E	G	G	G	D	D	D	C	C	H	H	G	D
TestO13	A	A	E	B	A	A	E	E	H	A	E	B	H	E	A	A	A	A	B	D
TestO14	C	C	G	G	D	B	C	C	C	C	G	G	G	C	B	B	B	B	C	G
TestO15	H	H	B	B	H	H	A	A	H	H	D	D	B	B	H	C	H	H	B	B
TestO16	D	D	G	B	D	D	E	E	D	D	D	D	D	D	B	B	D	D	B	D
TestO17	F	B	D	D	B	B	D	D	C	B	B	B	B	B	A	A	B	B	C	C

Those results are then divided into two separate tables as the followings that are based on the number of n to ease the analysis process. Table 6 shows the results of 1-grams experiments in all five algorithms.

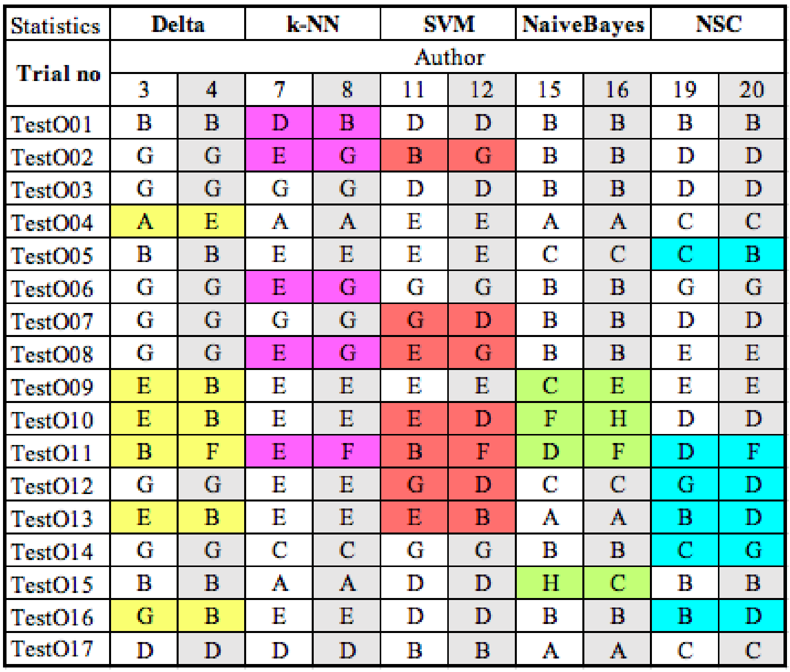
Table 6: The result of experiments done with n-grams = 1



Statistics	Delta		k-NN		SVM		NaiveBayes		NSC	
Trial no	Author									
	1	2	5	6	9	10	13	14	17	18
TestO01	G	G	A	A	G	G	A	C	G	G
TestO02	C	D	D	D	D	D	C	A	G	G
TestO03	G	G	C	C	G	G	A	G	C	C
TestO04	G	G	A	A	G	G	D	E	A	E
TestO05	C	C	A	A	C	C	A	E	E	E
TestO06	D	D	D	D	G	D	A	E	F	F
TestO07	A	A	D	D	A	A	A	A	D	F
TestO08	F	F	F	F	F	F	B	B	F	F
TestO09	E	E	E	E	E	E	B	E	E	E
TestO10	F	F	H	H	F	F	B	B	H	H
TestO11	F	G	D	D	G	G	B	E	C	F
TestO12	C	C	C	C	G	G	D	D	H	H
TestO13	A	A	A	A	H	A	H	E	A	A
TestO14	C	C	D	B	C	C	G	C	B	B
TestO15	H	H	H	H	H	H	B	B	H	H
TestO16	D	D	D	D	D	D	D	D	D	D
TestO17	F	B	B	B	C	B	B	B	B	B

As can be seen, several cells of the tables are in different colors showing the inconsistency of the outcome of authorship attribution. Overall, k-NN provides a result with the least inconsistency: 94% of the attribution is consistent. On the other hand, Delta, SVM, and NSC have the same number of consistency result –only three out of seventeen results that are inconsistent. Naïve Bayes still performs not very well looking at its number of inconsistency: ten out of seventeen results are inconsistent.

Table 7: The result of experiments done with n-grams = 3



Statistics	Delta		k-NN		SVM		NaiveBayes		NSC	
Trial no	Author									
	3	4	7	8	11	12	15	16	19	20
TestO01	B	B	D	B	D	D	B	B	B	B
TestO02	G	G	E	G	B	G	B	B	D	D
TestO03	G	G	G	G	D	D	B	B	D	D
TestO04	A	E	A	A	E	E	A	A	C	C
TestO05	B	B	E	E	E	E	C	C	C	B
TestO06	G	G	E	G	G	G	B	B	G	G
TestO07	G	G	G	G	G	D	B	B	D	D
TestO08	G	G	E	G	E	G	B	B	E	E
TestO09	E	B	E	E	E	E	C	E	E	E
TestO10	E	B	E	E	E	D	F	H	D	D
TestO11	B	F	E	F	B	F	D	F	D	F
TestO12	G	G	E	E	G	D	C	C	G	D
TestO13	E	B	E	E	E	B	A	A	B	D
TestO14	G	G	C	C	G	G	B	B	C	G
TestO15	B	B	A	A	D	D	H	C	B	B
TestO16	G	B	E	E	D	D	B	B	B	D
TestO17	D	D	D	D	B	B	A	A	C	C

There are more inconsistencies resulted from the 3-grams test as can be seen from the above table that more colors appear. It is interesting to see that when the n = 1

NaiveBayes performs not very well, however, it performs better with n = 1. It is evident that it has the least inconsistency among other algorithms –only four from the results are inconsistent. Following NaiveBayes, k-NN is the second algorithm that performs better compared to the other looking at its four inconsistencies of the test result. Delta performs as well as NSC with approximately 64% percent of consistency. The least consistent is SVM showing only ten of the results that are consistent.

Based on the result elaborated above, it can be concluded that k-NN provides better consistency when the n = 1 –it is 94% consistent, yet NaiveBayes provides better consistency when the n = 3 –it is 76% consistent. However, after quantifying the percentage of each algorithm both with the n = 1 and n = 3, subsequently counting the mean of each algorithm, it can be drawn that k-NN performs better in solving the task of authorship attribution open class with the percentage of 82% consistency. Thus, the result of the k-NN test with the n = 1 will be as the final result of the open class test since it has the higher percentage of consistency than the one with n = 3.

The comparison between these experiments' result and the actual answers can be seen in table 8. It is noticeable that only four attributions that are similar, they are testC09, testC15, testC16, and testC17. Hence, it can be concluded that this kind of approach may be not the best choice to perform in an open class task since the text is always assigned to one of the candidate authors whereas in reality a text may be written by none of the candidate authors.

Table 8: the comparison between the experiments' results with the actual answers

File Name	Results of Experiments	Actual Answers	File Name	Results of Experiments	Actual Answers
testO01	A	G	testO10	H	F
testO02	D	None	testO11	D	None
testO03	C	None	testO12	C	None
testO04	A	None	testO13	A	None
testO05	A	None	testO14	B	None
testO06	D	None	testO15	H	H
testO07	D	A	testO16	D	D
testO08	F	C	testO17	B	B
testO09	E	E			

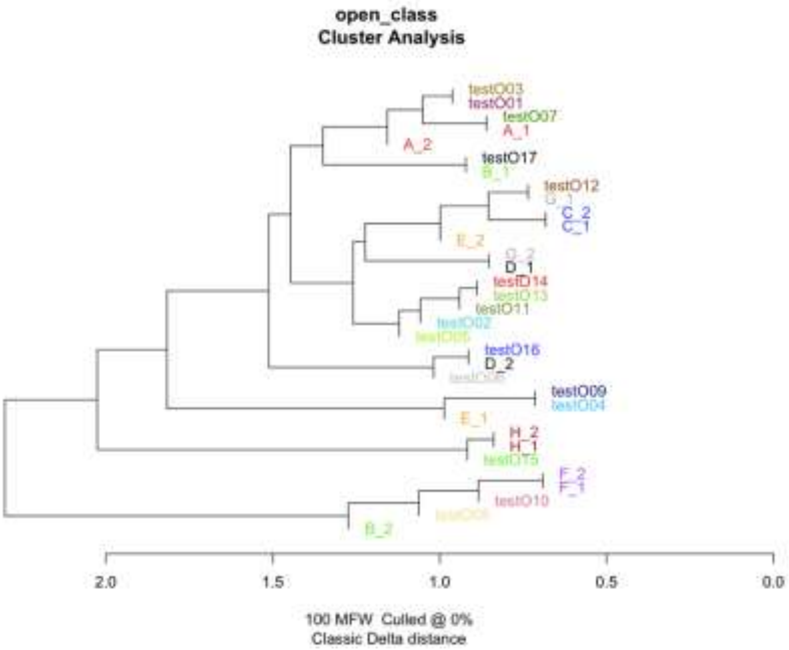


Figure 2: the result of stylo function in open class

As a comparison, cluster analysis using stylo function is done in open class task whose result can be seen in the following figure. Stylo function provides many different results compared to the experiments' results and actual answers. First of all, testO01 is assigned to author A based on the experiments and author G in actual answer, yet it is grouped testO01 in the same level branch with testO07 and one of A's writings. In addition, testO06 is grouped in the same level branch as testO16 and one of D's writing. These may be why the tool assigns testO01 to be written by author A in the experiment. Also, testO12 is clustered with one of G's writing in the same level branch with two C's writings that may lead the tool to assign C as the writer of testO12 in the experiments. Another striking difference is that testO02, testO05, testO11, testO13, and testO14 are assigned to none of the candidate authors –which are different from the actual answer that assigns to none the following text test: testO02 to testO06, and testO11 to testO14.

Interestingly, in cluster analysis testO04 and testO09 is grouped together in one branch and under the same level branch with one of the E's writings that may lead the experiments as well as actual answer to assign testO09 to be written by author E. However, it is different for testO04 that logically may be assigned to E based on the experiment, but it turns out not. TestO08 is assigned to author F based on the experiment and author C in the actual answer but looking at the cluster analysis result, testO08 is clustered together with testO10 and another branch of two F's writings. Logically, testO08 will be assigned to be written by F as F is the closest author in the tree diagram.

CONCLUSION

Based on the experiments of both open and closed class, the analysis of the results of both classes, the comparison of results between stylo and classify functions, and the comparison between experiments' result with the actual answer, this tool seems reliable for solving the authorship attribution for closed-class owing to the fact that it has provided a 100% consistent result using SVM algorithm. For the open class, the k-NN algorithm may be the best choice to perform the task as it reaches 94% of consistency. What needs to keep in mind is that those two results are using the 1-grams test. In comparison with the result of cluster analysis using stylo function, it can be seen that there are two different attributions compared to the actual answer. In this kind of situation, conventional qualitative analysis may be best to carry out to find out the differences and can define the author then. In the open class, it may be better to perform stylo than classify function since stylo function provide results closer to the actual answer while classify function results on a mere of four correct answers.

The relationship between the number of n and the consistency of the test result has not been explored yet in this research. In doing authorship attribution for the closed class, it is recommended to use Stylo tool for its consistency. However, it is not yet definitive that this tool is the right methodology to solve the task as comparing between two or more tools will provide more choices and more chances to explore to find the most appropriate one. Moreover, the legal system most of the time challenges authorship analysis as it does not have any valid methodology to perform an analysis but analyzing styles using stylometry and quantifying them with the help of computational method may make the analysis more sense to the legal system that hopefully leads to more informed decisions.

REFERENCES

De Vel, O., et al. (2001). Mining E-mail Content for Author Identification Forensics. *ACM Sigmod Record, 30(4), 55-64.*

Eder, M., Rybicki, J., and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *"R Journal", 8(1): 107-121.*

Grant, T. (2007). Quantifying Evidence in Forensic Authorship Analysis. *International Journal of Speech, Language & the Law, 14(1).*

Grant, T. (2008). Approaching Questions in Forensic Authorship Analysis. *Dimensions of Forensic Linguistics, 5, 215.*

Grant, T. (2010). Text Messaging Forensics. Txt 4n6: Idiolect Free Authorship Analysis?. In M. Coulthard and A. Johnson (Ed.), *The Routledge Handbook of Forensic Linguistics (pp. 508-522).* Abingdon: Routledge.

Grant, T. (2013). TXR 4N6: Method, Consistency, and Distinctiveness in the Analysis of SMS Text Messages. *Journal of Law and Policy, 21(2), 467-494.*

Gray, A., MacDonell, S., & Sallis, P. (1997). *Software Forensics: Extending Authorship Analysis Techniques to Computer Programs.* Retrieved December 20, 2018 from https://ourarchive.otago.ac.nz/handle/10523/872.

Hughes, V. (2013, July 19). How Forensic Linguistics Outed J.K. Rowling (Not to Mention James Madison, Barack Obama, and the Rest of Us). *National Geographic.* Retrieved

20 December 2018, from https://www.nationalgeographic.com/science/phenomena/2013/07/19/how-forensic-linguistics-outed-j-k-rowling-not-to-mention-james-madison-barack-obama-and-the-rest-of-us/.

Iqbal, F., et al. (2013). A Unified Data Mining Solution for Authorship Analysis in Anonymous Textual Communications. *Information Sciences, 231, 98-112.*

Jeffreys, B. (2018, December 14). Cheating university students face FBI-style crackdown. *BBC News.* Retrieved December 17, 2018, from https://www.bbc.com/news/education-46530639.

Juola, P. (2008). *Authorship Attribution. Foundations and Trends in Information Retrieval. Vol 1. n.3.* Boston: NOW Publishers.

Luyckx, K. (2010). *Scalability Issues in Authorship Attribution.* A Doctoral Dissertation. Antwerp University.

MacLeod, N., & Grant, T. (2012). Whose Tweet? Authorship analysis of micro-blogs and other short-form messages. In S. Tomblin, N. MacLeod, R. Sousa-Silva, & M. Coulthard (Eds.), *Proceedings of the International Association of Forensic Linguists' tenth biennial conference (pp. 210-224).* Aston University.

McMenamin, G.R. (2002). *Forensic Linguistics: Advances in Forensic Stylistics.* Boca Raton: CRC Press LLC.

Oliveira, B., van der Voet, J., and Jazilah, N. (2018). *Protocol for Authorship Analysis.*

Olsson, J. (2009). *Word Crime: Solving Crime through Forensic Linguistics.* London, Bloomsbury.

Peng, J., Choo, K. K. R., & Ashman, H. (2016). Bit-level N-gram Based Forensic Authorship Analysis on Social Media: Identifying Individuals from Linguistic Profiles. *Journal of Network and Computer Applications, 70, 171-182.*

Solan, L.M. (2010). The Forensic Linguist: The Expert Linguist Meets the Adversarial System. In M. Coulthard and A. Johnson (Ed.), *The Routledge Handbook of Forensic Linguistics (pp. 395-407).* Abingdon: Routledge.

Solan, L.M. (2013). Intuition versus Algorithm: The Case of Forensic Authorship Attribution. *Journal of Law and Policy, 21(2), 551-576.*

Verhoeven, B. (2015, May). *Computational Stylometry.* Guest Lecture at Universite Libre de Bruxelles. Retrieved May 20, 2018 from https://pdfs.semanticscholar.org/. presentation/e9ab/e5010a5ba3c71dac08ab9f43ec38fce66906.pdf

Zheng, R., et al. (2003, June). Authorship Analysis in Cybercrime Investigation. In *International Conference on Intelligence and Security Informatics (Pp. 59-73).* Springer, Berlin, Heidelberg.