IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 14, 2022, accepted January 30, 2022, date of publication February 16, 2022, date of current version March 1, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3152201

Text Mining in 19th-Century Essays for Investigating a Possible Collaborative Authorship Problem: John Stuart Mill and Harriet Taylor Mill

ANDREAS NEOCLEOUS[1], **GIORGOS KATALIAKOS**[2], **AND ANTIS LOIZIDES**[2]
[1]Department of Computer Science, University of Cyprus, Aglantzia, 2109 Nicosia, Cyprus
[2]Department of Social and Political Sciences, University of Cyprus, Aglantzia, 2109 Nicosia, Cyprus

Corresponding author: Andreas Neocleous (neocleous.andreas@gmail.com)

This work was supported by the University of Cyprus through a starting-up grant.

ABSTRACT In this work, we use machine learning techniques to address a research question regarding the authorship of two famous essays in the nineteenth century. *On Liberty* (1859) and *The Subjection of Women* (1869) were published under John Stuart Mill's name, a widely studied nineteenth-century British philosopher. Mill himself attributed them to collaboration with his wife and partner, Harriet Taylor Mill. More than 150 years later, the question remains whether the author of these two canonical texts in the history of political thought was solely John Stuart Mill. Experts are divided on taking John Stuart Mill's attribution at face value, since Harriet Taylor Mill had died in 1858. Addressing this question, we use a dataset consisted in essays of both authors, to train three state-of-the-art classifiers that are able to learn and distinguish the writing style of each author. Then, we use the models built to attribute the two famous essays of disputed authorship to one of the two. From the results, we conclude that the classifiers are able to learn the two classes very well, and they return high accuracies on the validation set. Regarding the test set, most of the models attribute the two essays to John Stuart Mill, however, the contribution of Harriet Taylor Mill is shown for some chunks of text of both essays. These results, we conclude, explain why experts are divided on this particular research question.

INDEX TERMS Authorship attribution, text classification, machine learning, feature selection.

I. INTRODUCTION

Computer-based or computer-assisted authorship identification tries to answer an old question with new means: who is the real author of a piece of writing. Traditionally, researchers have used textual and contextual evidence to verify or to dispute the authorship of a text. By late-nineteenth and early-twentieth century, literary investigations had incorporated statistics-related techniques. A century later, the availability of intelligent tools suitable to the task has been expanded dramatically [2]–[5].

In recent years, adapting to specific research questions, scholars have proposed a number of advanced and automated authorship identification methods, that lie within the area of text mining. These methods go beyond statistics, while making use of machine learning and artificial intelligence methods including deep learning (DL). The need for using DL methods in text mining was raised by the big amount of text and information that became available mainly from the internet, e.g. social media and blogs. It is well known that DL methods work better when trained with many examples and therefore for the task at hand we think that it is not the proper method to use, simply because our data are not enough in volume or size. One example of such application is presented in [10]. In this work a dataset of 50M tweets is used for training which makes the decision of using DL reasonable.

In this paper we build on our previous work [1]. There made our first attempt to use standard machine learning techniques for creating models that can distinguish the writing style of three different authors in three different classes.

The associate editor coordinating the review of this manuscript and approving it for publication was Biju Issac.

At the same time, we created another class in those models to distinguish essays known to be the product of collaboration.

John Stuart Mill (1806-1873) was one of the most influential British writers in the nineteenth century. He was known to be the sole author of *On Liberty* and *The Subjection of Women*. However, in his *Autobiography* (1873), Mill openly shared authorship with his wife, Harriet Taylor Mill (neé Hardy), in writing the first and with both Harriet Taylor Mill and her daughter, Helen Taylor, in writing the second. Mill's contemporaries were not convinced about the attribution (e.g. [14]–[16]); neither were most scholars who examined the issue in mid-twentieth century (e.g. [17]–[20]), despite some voices accepting Mill's claims ([21], [22]). More recently, the reluctance to accept Mill's sharing authorship credit with two women, either in the nineteenth or the twentieth century, was associated to a general reluctance to acknowledge women's role in the history of philosophy ([26]–[28]).

In short, we begin from the premise that, given the importance of *On Liberty* and *The Subjection of Women* for the history of political thought, the divergence of opinion as regards John Stuart Mill being the sole author is reason enough to attempt to tackle the issue with non-traditional methods. It is important to get the author(s) of these texts right. Of course, in this paper we study authorial patterns, not philosophical influence. Still, for Harriet Taylor Mill herself, the two are closely related: "It would perhaps not be possible to find two minds accustomed to think for themselves whose thoughts on any identical subject should take in their expression the same form of words" (Taylor Mill, 1998: 140).

As just mentioned, there are three well-defined approaches to the question of Harriet Taylor Mill's role in the writing process of these texts. In a sort of backhanded compliment, the first acknowledges Harriet Taylor Mill's role in those works (and some other) by John Stuart Mill. We call it a backhanded compliment because the evidence offered is a lack of clear and consistent reasoning in the texts [21]–[23]. The second dismiss Harriet Taylor Mill's philosophical role in any important work which bore John Stuart Mill's name. This means, on the one hand, that John Stuart Mill's "important works" might bear traces of Harriet Taylor Mill's stylistic influence, while, on the other hand, what counts as an "important text" varies ([17]–[20], [24]). For this group of scholars, Harriet Taylor Mill's influence increases as the importance of the work lessens. The third group of scholars, try to break the above dichotomous mold, by taking John Stuart Mill at his word when he assigned Harriet Taylor Mill co-author status to both *On Liberty* and *The Subjection of Women* (even though these works were published bearing solely his name and there is no traditional evidence to support this claim, i.e. manuscripts of these texts featuring her corrections, suggestions, or additions). They argue that Harriet Taylor Mill was a highly original, radical, and accomplished woman philosopher and her contributions to those two works must be recognized to be more than just stylistic ([25]–[28]). As we shall see, our results confirm stylistic similarity to both, even though the bulk of the text is attributed to John Stuart Mill. Who wrote what part in a collaborative writing is a notoriously difficult problem to solve (e.g., compare [29] with [30]).

From the results in [1], we concluded that creating a four-class classification model is not the optimal way with which to go through. In this paper, we simplify the models into training binary classifiers to distinguish between two authors: John Stuart Mill and Harriet Taylor Mill. We divide every essay into chunks of specific length of words and we use those chunks as separated instances, instead of feeding the entire essay as a single instance. This change in methodology has several benefits. First, we increase the size of the training set. Second, we use the same size for every instance, which can be considered as one of the normalization stages. Third, we gain a better understanding with a more fine-grained analysis of every essay. For example, no only can we have an attribution result for the entire essay, but also for parts of them. From this, we can observe and make assumptions for specific parts of contribution by other authors in an essay. In the process, we also examine whether the conclusion that it was John Stuart Mill who wrote the two texts in question still holds, despite the change in methodology.

II. PREVIOUS WORK

In recent years, adapting to specific settings and research questions, scholars have proposed a number of automated authorship identification methods. In the most basic formulation of the authorship identification task, researchers examine whether the same person authored two, possibly short texts. Moshe Koppel *et al.* [31] argue that if we can solve this "fundamental problem", we can solve any other authorship identification problem either in ideal settings or in less-than-ideal or in real-life settings. There are a number of factors that guide the selection of feature extraction at any given authorship identification task – e.g., language, medium, topic, genre. Broadly speaking, the number of candidate authors and the length of text available (both eponymous and anonymous) more or less guide the choice of method in the quest for better results. For the sceptic, however, what's better might never be good enough to replace the human expert. (see further, [32]).

First, let's take a look at authors: the number of candidates who claim (or are assigned) authorship for a text. In a closed set of candidate authors, we are certain that one author from the set – small or large – is the author of the anonymous work. In style variation studies (see e.g., [33], [34]), the closed set is populated by just one person. For example, already by mid-to-late 19th century scholars have drawn on stylometric data to make a case on the chronology of Plato's dialogues (e.g., [35], [36]; see further, [37]). Variation in authorial style complicates the authorship identification task, though not enough to invalidate its results [38].

Second, let's turn to texts. Researchers use a variety of features, often in conjunction, to make claims as regards authorial fingerprints (see further, [4], [5], [32]). In 2009, Efstathios Stamatatos [4] grouped a number of studies together with reference to the stylometric features they used to define authorial uniqueness, mostly syntactic, lexical, semantic, and character features. Syntactic feature extraction looks at sentence structure, as the syntactic patterns (punctuation, function words, etc.) seem to be a reliable marker for authorship. Lexical feature extraction focuses on words: once the text is tokenized, the researcher may focus on word – and sentence – length, word frequencies and word sequences, errors as well as vocabulary richness. Semantic feature extraction looks at the meaning of words or sentences: how one handles synonym pairs or how the appearance of certain words depends on other words can be reliable markers of authorial identity under certain conditions. Character feature extraction does not stop at the level of words or sentences, but deals with characters – e.g. letters (lower- and/or uppercase), digits, punctuation marks. There are other application-specific ways to mine for relevant information in a text, but character- and lexical-feature extraction are the most popular [4].

Another issue to consider when we are looking at texts is the minimum corpus size required to identify uniqueness. Burrows suggests that a 10000-word corpus size is reliable enough to construct an author's set (with a minimum of 500 words for each text sample included in the set) [6]. Maciej Eder cuts Burrows's number to 5000 words per authorial set, pointing out that, within such an authorial set, smaller chunks of single texts, well below 500 words, still produce reliable results, even in cross-language settings [7]. Eder however warns that even though a lower threshold is possible (Eder 2017 ran successful tests at 2000 words), other factors might distort the authorial signal. Topic and genre are such distorting factors, even though cross-topic attribution/verification is considered a less complex problem than cross-genre attribution/verification ([8], [9]).

Third, we move on method. There are two families of methods used in authorship identification studies [4]. Profile-based methods, as in [4] terms them, try to create a cumulative representative style of the known author's text, which is then compared with the style of the test corpus. Authors are ranked according to the distance which keeps their style from the style of the anonymous text farther away. Nuances in style between different works of the known author are disregarded, since all the eponymous works used are grouped into one large document during feature extraction. The second family of methods, that of instance-based methods [4], uses separate eponymous texts to train a classifier, an algorithm which learns how input data relate within a class. The classifier is then employed to sort the unknown text under the appropriate author (whether author A or B or C in a small, closed-set of authors at an authorship attribution task or author A or Not-A in the two-class classification rendition of the authorship verification task).

III. METHODS

A. OVERVIEW

The pipeline of our system is shown in Fig. 1. The initial dataset consists of $D_n = 30$ texts. It includes 10 chapters from two essays by John Stuart Mill (total length of 55,862 words), 11 essays by Harriet Taylor Mill (total length of 15,082 words) that form the training and validation sets and nine chapters from two essays (*On Liberty* and *The Subjection of Women*) of an unknown author (total length of 90,858 words) that form the test set. The first step is to concatenate all the texts of every class into one long text, and then use it to create a series of instances of a pre-defined number of words. This technique reminds a frame based approach of a time series. The only difference here is that we do not apply any overlap between consecutive instances, i.e. an instance starts one word after the last word of the previous instance. In this way, we create a dataset that every instance has the same length, but most importantly, we increase dramatically the number of instances, which allows the classifiers to be trained better. For example, in our dataset we start from 30 texts, and with an instance length of 100 words we create a dataset of about 900 instances.

After this is done, we use the above mentioned dataset to split it into two smaller datasets which form the "training set" (D_{tr}) and the "validation set" (D_{val}). The training set is chosen randomly to form the 70% of the initial dataset and the rest is kept for validation. The test set (D_{test}) remains always the same. Next, we use the training set to extract the training feature set (f_{tr}) and from that to extract the same features to define the validation (f_{val}) and the test feature sets (f_{test}).

Considering that the n-grams consist of all the unique words and their unique pairs of the training set, this creates a very large feature set. This might generate several issues when trying to train a classifier, such as overfitting or overuse of computational power. To overcome this, we perform a dimensionality reduction technique, the widely used technique called "Principal Component Analysis (PCA)". The new reduced feature set is shown in Fig. 1 as (f'_{tr}). Using the model of the PCA, we also compute the f'_{val} and f'_{test}.

In this study we create different feature sets to train three classifiers. One feature set for example is to use only the punctuations, one other is to use only the unigrams, etc. In Section III-C we elaborate on that. After we train the classifiers and get the models, we present the validation and test sets to get the results. Finally, we repeat this procedure three times by using different training and validation sets (three-fold cross validation) in order to make sure that our results are correct and robust.

For every fold we compute the accuracies for both training and test sets. The results of the test set are reported using statistics.

B. DATA

1) OVERVIEW

In Table 1 we present the details of our dataset. We have collected 13 essays (first column in Table 1) by two authors to

TABLE 1. The dataset used for training and test.

Essay #	Author	Essay name	Year	Number of words	Training or Test
1	John Stuart Mill	Utiliarism (Ch. 1)	1861	1857	Training
1	John Stuart Mill	Utiliarism (Ch. 2)	1861	8552	Training
1	John Stuart Mill	Utiliarism (Ch. 3)	1861	3497	Training
1	John Stuart Mill	Utiliarism (Ch. 4)	1861	2898	Training
1	John Stuart Mill	Utiliarism (Ch. 5)	1861	9796	Training
2	John Stuart Mill	Considerations on Representative Government (Ch. 1)	1861	4374	Training
2	John Stuart Mill	Considerations on Representative Government (Ch. 2)	1861	7586	Training
2	John Stuart Mill	Considerations on Representative Government (Ch. 3)	1861	6923	Training
2	John Stuart Mill	Considerations on Representative Government (Ch. 4)	1861	4124	Training
2	John Stuart Mill	Considerations on Representative Government (Ch. 5)	1861	6091	Training
3	Harriet Taylor Mill	Australia	1831	241	Training
4	Harriet Taylor Mill	German Prince	1832	225	Training
5	Harriet Taylor Mill	Manners	1832	1331	Training
6	Harriet Taylor Mill	Hampden	1832	2635	Training
7	Harriet Taylor Mill	Mirabeau	1832	1416	Training
8	Harriet Taylor Mill	Plato	1832	623	Training
9	Harriet Taylor Mill	French Revolution	1832	2536	Training
10	Harriet Taylor Mill	Seasons	1832	1613	Training
11	Harriet Taylor Mill	Conformity	MS c1831	1934	Training
12	Harriet Taylor Mill	Laconicisms	MS c1832	1652	Training
13	Harriet Taylor Mill	Alroy	MS c1833	601	Training
14	Unknown	The Subjection of Women (Ch. 1)	1869	12169	Test
14	Unknown	The Subjection of Women (Ch. 2)	1869	8868	Test
14	Unknown	The Subjection of Women (Ch. 3)	1869	12717	Test
14	Unknown	The Subjection of Women (Ch. 4)	1869	9963	Test
15	Unknown	On Liberty (Ch. 1)	1859	5607	Test
15	Unknown	On Liberty (Ch. 2)	1859	16024	Test
15	Unknown	On Liberty (Ch. 3)	1859	7830	Test
15	Unknown	On Liberty (Ch. 4)	1859	7881	Test
15	Unknown	On Liberty (Ch. 5)	1859	9354	Test

train and validate the system: two essays by John Stuart Mill that are split in five chapters each and 11 by Harriet Taylor Mill. The test set consists of two essays (*On Liberty* and *The Subjection of Women*) of disputed authorship. In the second and third columns of Table 1 we present the author and the titles of the essays used in our dataset. In the fourth column we provide the year that the essay was written and in the fifth column we provide the number of words. In the last column, we mark every essay with the tags "Training" or "Test" to indicate how they were treated in the modelling procedure.

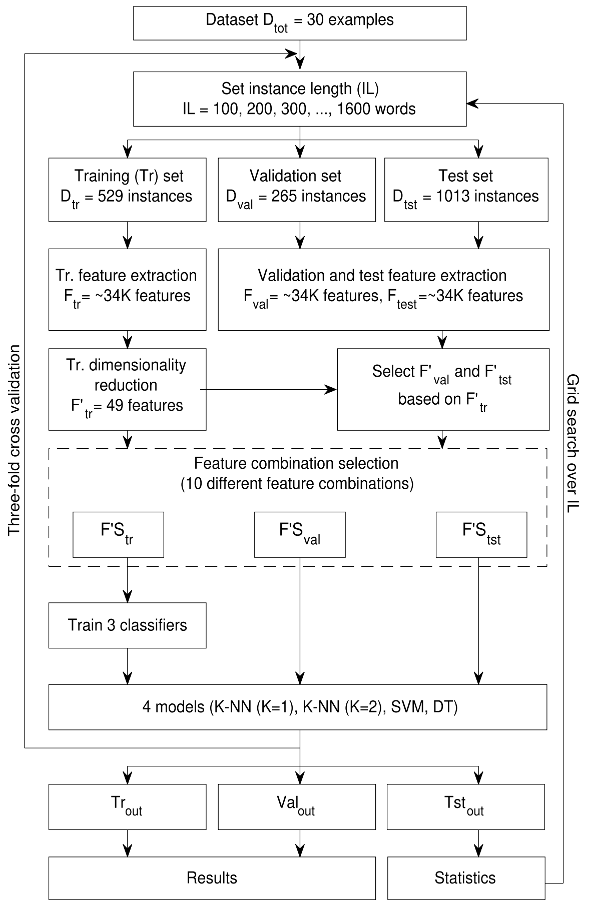
2) DEVIATIONS FROM PREVIOUS DATASET

Following the corpus in [1], we used the standard editions for John Stuart Mill's [39] and Harriet Taylor Mill's [40] authorial set. It should be noted that, unlike with John Stuart Mill, there was no much choice of texts in the case of Harriet Taylor Mill. In both cases, we used texts with little evidence of interference by the other (e.g. when part of the manuscript was in Mill's hand, or when Mill assigned the piece as joint authorship). In the case of Harriet Taylor Mill, this decision eliminated *Remarks on Mr. Fitzroy's Bill for the More Effectual Prevention of Assaults on Women and Children* (1853) and *The Enfranchisement of Women* (1851). Eliminating *The Enfranchisement of Women* (1851) was a tough choice to make, having been attributed solely to Harriet Taylor Mill a few years after its publication. It was originally published anonymously and some thought it was by John Stuart Mill. His role, as he tried to explain, was limited to serving as interlocutor, amanuensis and copy-editor to his wife in the process of writing. A contemporary critic, however, thought that this essay was a poor imitation, a parody, of John Stuart Mill's style [16]. To avoid thus the possibility of contamination of Harriet Taylor Mill's training set by John Stuart Mill's authorial style we exclude these two essays from the training procedure, even though they would almost double its size. Also, texts by Helen Taylor are not included, both to simplify the models and to focus on the collaboration between John Stuart Mill and Harriet Taylor Mill.

C. FEATURE DEFINITION

The features we use for the task at hand are separated in ten main categories: 1) The "Counts" category which includes statistical properties (average and standard deviations) of "sentence length" and "word length", 2) the "Punctuations" category which consists of 17 selected punctuations, 3) the "CLAWS tags" category (the Constituent Likelihood Automatic Word-tagging System) which is a list of 138 grammatical pre-defined tags that are extracted using a tool that classifies every word in one of those 138 tags.[1] This tool is developed by the University Centre for Computer Corpus Research on Language (UCREL) and it is freely available online,[2] the "n-grams" category which consists of 4) Unigrams and 5) Bigrams, 6) all the features group together (1-5) and 7-10) the PCA of the categories 3-6. The unigrams describe the frequency of every word in a text, e.g., how many

[1] http://ucrel.lancs.ac.uk/claws7tags.html
[2] http://ucrel.lancs.ac.uk/claws/



Dataset D_{tot} = 30 examples
Set instance length (IL)
IL = 100, 200, 300, ..., 1600 words
Training (Tr) set
D_{tr} = 529 instances
Validation set
D_{val} = 265 instances
Test set
D_{tst} = 1013 instances
Tr. feature extraction
F_{tr}= ~34K features
Validation and test feature extraction
F_{val}= ~34K features, F_{test}=~34K features
Tr. dimensionality reduction
F'_{tr}= 49 features
Select F'_{val} and F'_{tst} based on F'_{tr}
Feature combination selection
(10 different feature combinations)
$F'S_{tr}$
$F'S_{val}$
$F'S_{tst}$
Train 3 classifiers
4 models (K-NN (K=1), K-NN (K=2), SVM, DT)
Tr_{out}
Val_{out}
Tst_{out}
Results
Statistics
Three-fold cross validation
Grid search over IL

FIGURE 1. The pipeline of our system. We perform a grid search over the parameter "instance length (IL)" while we test our models by applying a three fold cross validation. The methodology consists of, feature extraction, dimensionality reduction, training, validation, feature selection and results. The numbers for the feature size are for $IL = 100$ words.

TABLE 2. The different feature groups and their sizes that are used as inputs to train the classifiers.

#	Feature group name	Feature group size
1	All features	33,841
2	Counts	5
3	Punctuations	12
4	CLAWS tags	138
5	Unigrams	5782
6	Bigrams	27904
7	All features - PCA	49
8	CLAWS - PCA	52
9	Unigrams - PCA	4
10	Bigrams - PCA	4

times a word appears in a document. The bigrams describe the frequency of every consecutive pair of words in a document.

In this work we focus on testing several feature sets to identify the optimal set that can better distinguish the two requested classes. In the second row of Table 2 we present the names of the feature sets and in the third column the feature group size. It is shown that the initial entire feature sets consists of 33,841 features and their PCA consists of only 49.

D. STATISTICAL ANALYSIS

In this Section we present an overview of the statistical properties of the extracted dataset. We first examine the sensitivity of the parameter "instance length (IL)" that is used for a grid search. In Fig. 2 we present an illustration of the first two principal components of the CLAWS tags, for $IL = 100, 200, 300, \ldots, 1600$. We use scatter plots to visualize the training set of the two classes, together with the test set (black circles for John Stuart Mill and red circles for Harriet Taylor Mill and cyan dots for the test set). In this figure, it is clear that the discriminative ability of the selected features, in relation to the instance length increases. The two clusters discriminate better with $IL > 400$ words.

In Fig. 3, we present six scatter sub-plots of the three feature groups in categories 4–6 (see Chapter III-C and Table 2). The first column of sub-plots (Figs. 3a and b) show the distribution of the first two principal components of the entire feature set (group 4) for the training (above) and the validation and test sets (below). The second (Figs. 3c and d) and third (Figs. 3e and f) columns show the distributions of the "unigrams PCA" and "bigrams PCA" respectively, for $IL = 900$ words. In all of those figures we observe that the two classes separate nicely and the test set lies more on the class of John Stuart Mill.

E. CLASSIFIERS AND MODELLING

The question raised in this research can be approached by utilizing supervised learning methods. These methods use input vectors that are consisted by a number of features. Each one of those feature vectors represent one instance, which in this case is a number of consecutive words of specific length. Then, every instance is annotated with a label representing the class that belongs in.

In this paper we present the results of three classifiers: a) k-nearest neighbours (k-nn) for $k = 1$ and $k = 2$, support vector machines (SVMs) and decision trees. For further reading on the classifiers used, we refer to the following studies: Cover and Hart, [41], Cortes and Vapnik, [42], Durgesh and Lekha [43], Cervantes *et al.* [44], Tong and Koller [45], Wei *et al.* [46] and Rokach *et al.* [47].

F. COMPARISON WITH A BENCHMARK DATASET

In order to test the performance of our system, we used a benchmark dataset and we run the same experiments as explained above. We choose to use the dataset from a famous authorship attribution problem: "The Federalists papers". This dataset consists of a collection of 85 articles and essays written by Alexander Hamilton, James Madison, and John Jay under a single pseudonym.

In short, traditional methods of authorship attribution (e.g. biographical and historical) assigned all but twelve articles to their respective authors. The authorship of these twelve papers were claimed by both Hamilton and Madison. In mid-twentieth century, scholars began using non-traditional methods to settle the dispute between historians. The scholarly

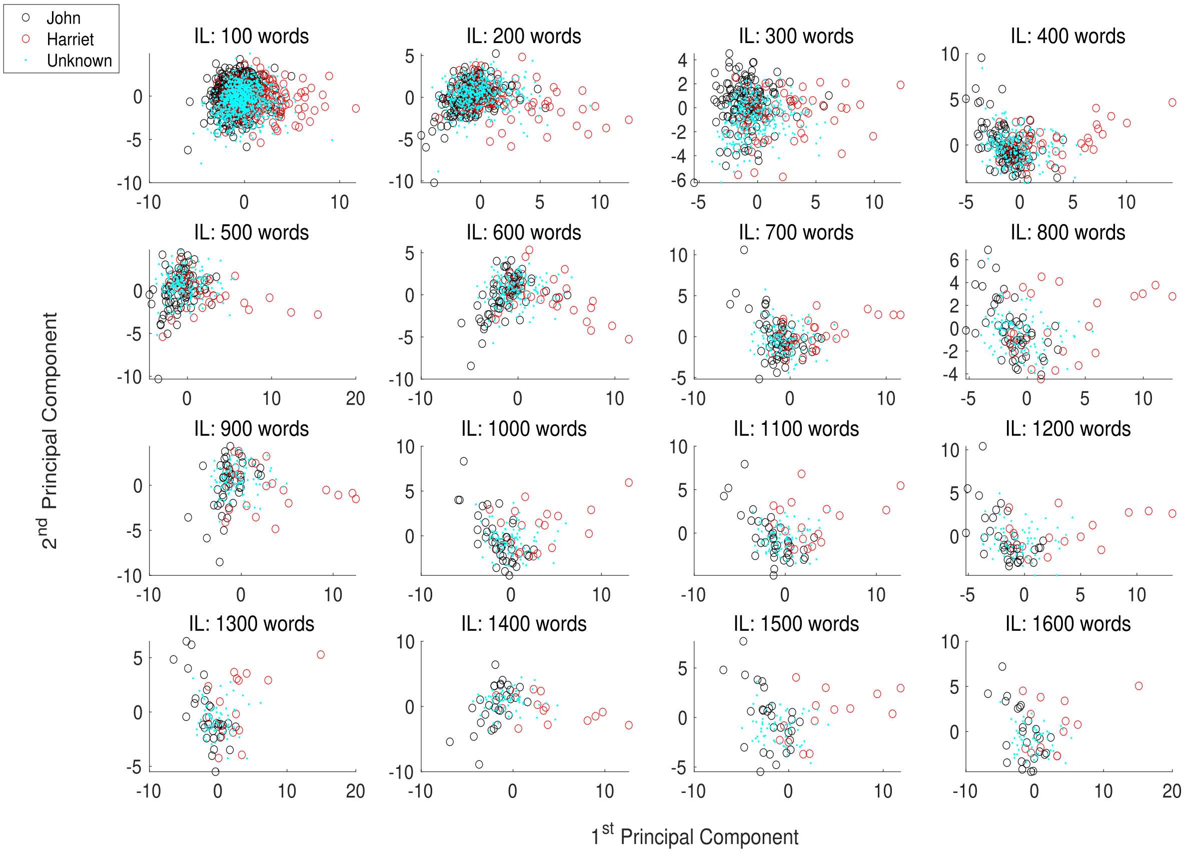


FIGURE 2. Scatter plots of the two training classes and the unknown class, using the first two principal components of the CLAWS tags. The 16 sub-plots show the distribution of the three classes in relation to the IL. The discriminative ability of those two features increases when the IL increases.

consensus is now that Madison was the author of those disputed articles, though admittedly on a couple of occasions the attribution gets tricky given the similarity of style and probable collaboration between the two. [48]–[55]

Our results return 97% of accuracy on the validation set and the disputed essays are attributed by all methods and all feature sets to Madison. These results suggest that our system works correctly.

IV. RESULTS

A. OVERVIEW

In this work we use essays of known authorship to train three classifiers. For examining the strength of several features, we build models using different feature groups and finally to examine the sensitivity of the instance length, we perform a grid search for lengths 100 to 1600 words, with a step of 100 words. To make sure that the results of the classifiers are robust and consistent, we perform a three-fold cross validation. The results of the three folds are consistent and therefore we can conclude that the dataset we use is robust. Here, we present the results of fold 1.

B. EVALUATION METRICS

In this work we choose to use the accuracy and the Matthews correlation coefficient (MCC) for evaluation metrics. The accuracy is a well known and very common metric in machine learning and it is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP, TN, FP and FN stand for true positive, true negative, false positive and false negative respectively. A true positive is considered when an instance is a chunk from John Stuart Mill and the method correctly classifies it as the one from John Stuart Mill. A true negative is considered when an instance is a chunk from Harriet Taylor Mill and the method correctly classifies it as the one from Harriet Taylor Mill. A false positive is considered when an instance is a chunk from Harriet Taylor Mill and the method incorrectly classifies it as the one from John Stuart Mill. A false negative is considered when an instance is a chunk from John Stuart Mill and the method incorrectly classifies it as the one from Harriet Taylor Mill.

We choose to also use the MCC metric which is mainly used when the two classes are imbalanced in size. The MCC ranges from -1 to $+1$, where ± 1 indicates perfect agreement or disagreement, and 0 indicates no relationship:

$$MCC = \frac{TPXTN - FPXFN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2}$$

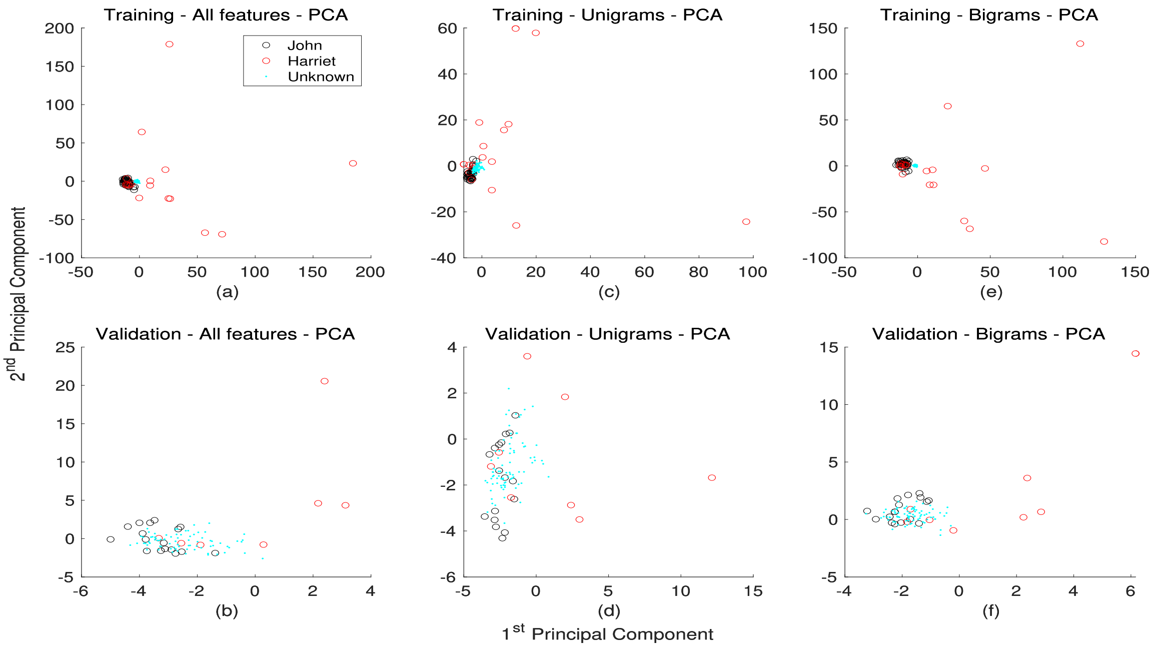


FIGURE 3. Scatter plots of the two training, validation and the unknown classes, using the first two principal components of three feature groups. The two classes separate nicely and the unknown class lies clearly on the class of John Stuart Mill.

C. TRAINING SET

All of the classifiers are able to learn the training set at a very good level for most of the models built, with accuracies of more than 90%. One example on these results is shown in Fig. 4, illustrating the accuracies of each classifier for $IL = 100, 200, 300, \ldots, 1600$, having as training set the entire feature set. In this figure we superimpose the training results with black dotted line together with the validation results in red dotted line. It is shown that K-NNs with $K = 2$ are not consistent on learning the training set.

D. VALIDATION SET

From Figure 4, we observe that the most consistent results between the training and the validation sets are the ones from SVM. It can be seen that the validation set returns accuracy of 100%, for ILs equal to 1200 words. For the case of the KNNs, the results between training and validation are consistent for $K = 1$ but there might be a possible overfitting in the training phase for $K = 2$. This is because the validation set is not classified correctly and in many cases the majority of the instances are misclassified in one of the two classes. The best validation results of the DTs are achieved with $IL = 800$ words.

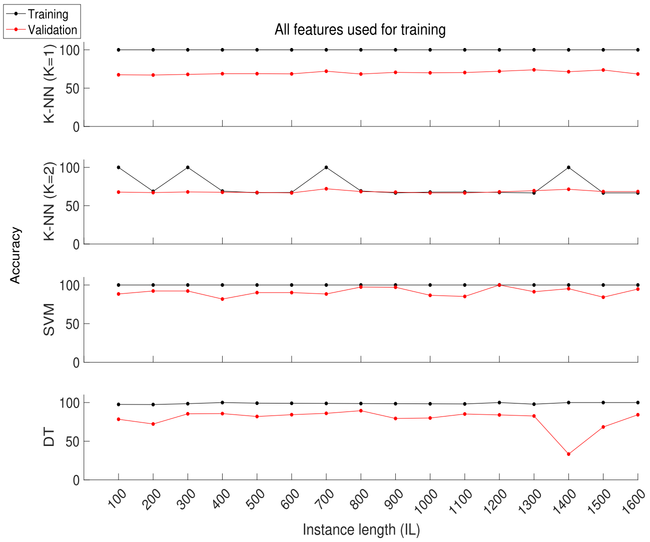


FIGURE 4. Training and validation results of all the classifiers used. The four sub-plots illustrate the accuracies of each classifier for $IL = 100, 200, 300, \ldots, 1600$, having as training set the entire feature set.

In Fig. 5 we present the results in accuracies of the classifiers used, for all the feature groups and for the IL that yields the best results on the entire feature set. The best results are achieved by the SVMs using the entire feature set and the "unigrams". High accuracy is also achieved with the bigrams. Regarding the dimensionality reduction, it is shown here that the results drop significantly. For example, the accuracy of the PCAs of the Unigrams drops at about 20%. The DTs return relatively high accuracies (>60%) for all the feature groups except the bigrams.

Another interesting finding here is the results achieved using the "counts" and the "punctuations" categories which are consisted with 5 and 12 features respectively. These features are considered by the community to have weak discriminative ability. Further on that, their size is impressively small to let the classifiers to be able to learn such highly non-linear feature spaces.

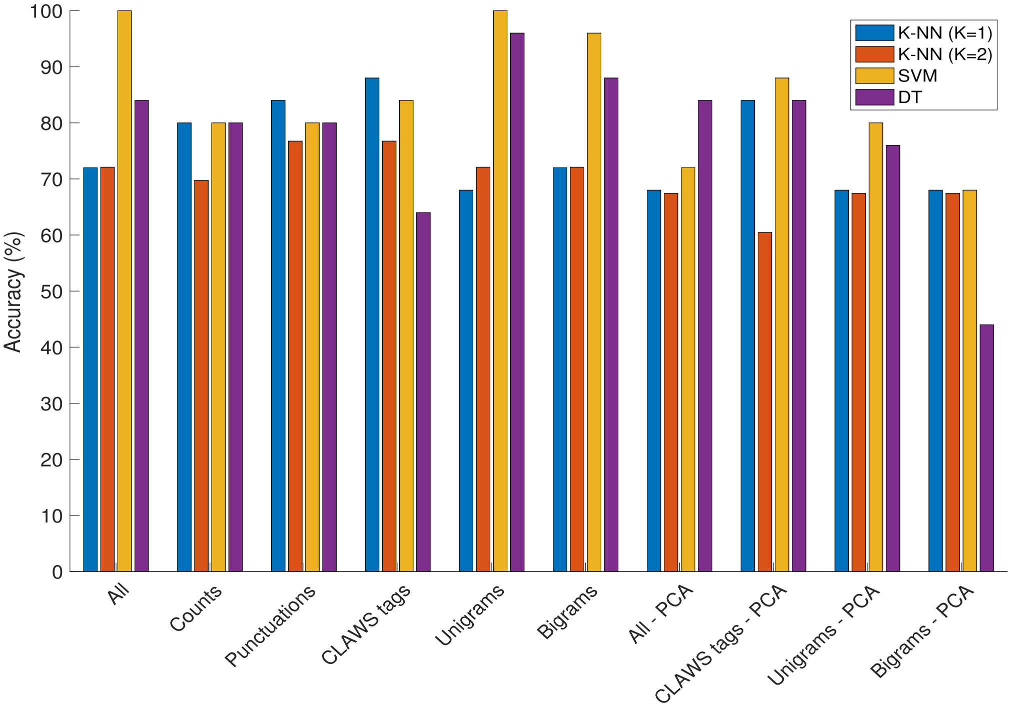


FIGURE 5. The accuracies of the validation set of the four models built for all the feature groups and for the parameter IL which yields the best results.

Overall, the SVMs seem to be the most consistent regarding the relation between training - validation results and they return 100% accuracies for the entire feature set and the ‘‘unigrams’’.

In Fig. 6 we present the results of the validation set in terms of the MCC evaluation metric. The first observation is that the K-NNs using the PCAs of the entire feature set, the ‘‘unigrams’’ and the ‘‘bigrams’’ suggest a random classification (MCC = 0). However for the PCAs of the CLAWS tags, the MCC is relatively high.

E. TEST SET

In Fig. 7 we illustrate the results on two essays (9 chapters in total) which are of an unknown / disputed author and they form the test set. These results are reported as a percentage of the instances that are attributed into one class divided by the total number of instances. The blue bars in Fig. 7 represent the class of John Stuart Mill while the red bars the class of Harriet Taylor Mill. The first impression here is that the majority of the validations of KNNs attribute the unknown essays to John Stuart Mill. However the SVMs attribute parts of some of the chapters of ‘‘SoW’’ and ‘‘OL’’ to Harriet Taylor Mill. These results are in line with [1].

V. DISCUSSION

In this paper, we are modelling the writing style of two nineteenth-century authors. We build models by feeding the input space with chunks of text of specific length of words. This is to create instances that are equal in size, but also to separate an essay in many small phrases, relating the whole system to act as a time signal. To test the sensitivity of this system, we applied a grid search by using different instance sizes and we observed that the classifiers learn the two classes better in bigger sizes.

The aim here is to use these models to classify two essays, whose authorship is being disputed. However, it is highly probable that the author is one of two authors, John Stuart Mill or Harriet Taylor Mill. The harder it is to distinguish between the other things being equal, the more likely the texts are products of collaboration. Given that Harriet Taylor Mill had died in 1858, a year before the publication of *On Liberty* and more than a decade before the publication of *The Subjection of Women*, it is surprising that it proved a very difficult task, not that the texts were attributed to John Stuart Mill. Given our models faired well in the Benchmark Dataset comparison, we take this result to be indirect evidence of collaboration.

The task itself is an interesting case for a machine learning point of view because not only were John Stuart Mill and Harriet Taylor Mill married; but also, they frequently exchanged views and collaborated on various writing projects (see part of the corpus labelled ‘‘joint productions’’ in [1]) since early 1830s. This is most likely the reason why it is so difficult to set the writing style of these authors apart. While there is contextual evidence (e.g. biographical and historical) to confirm that they shared ideas, and sheets of paper, there is no traditional way to verify that Harriet Taylor Mill guided John Stuart Mill's hand as he wrote *On Liberty* and *The Subjection*

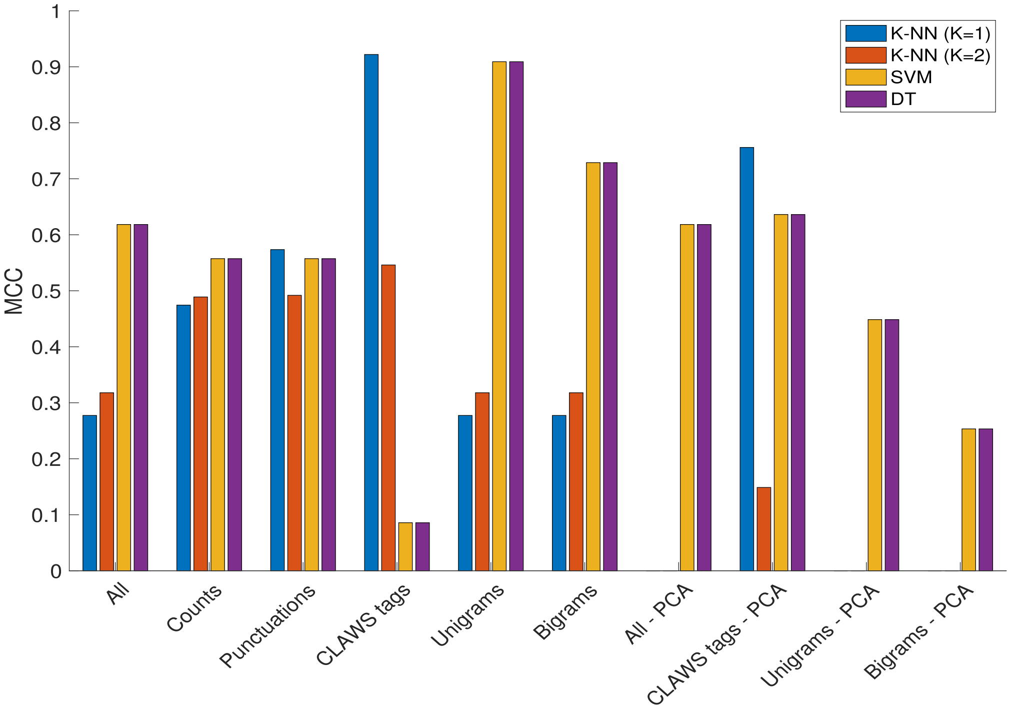


FIGURE 6. The MCCs of the validation set of the four models built for all the feature groups and for the parameter IL which yields the best results.

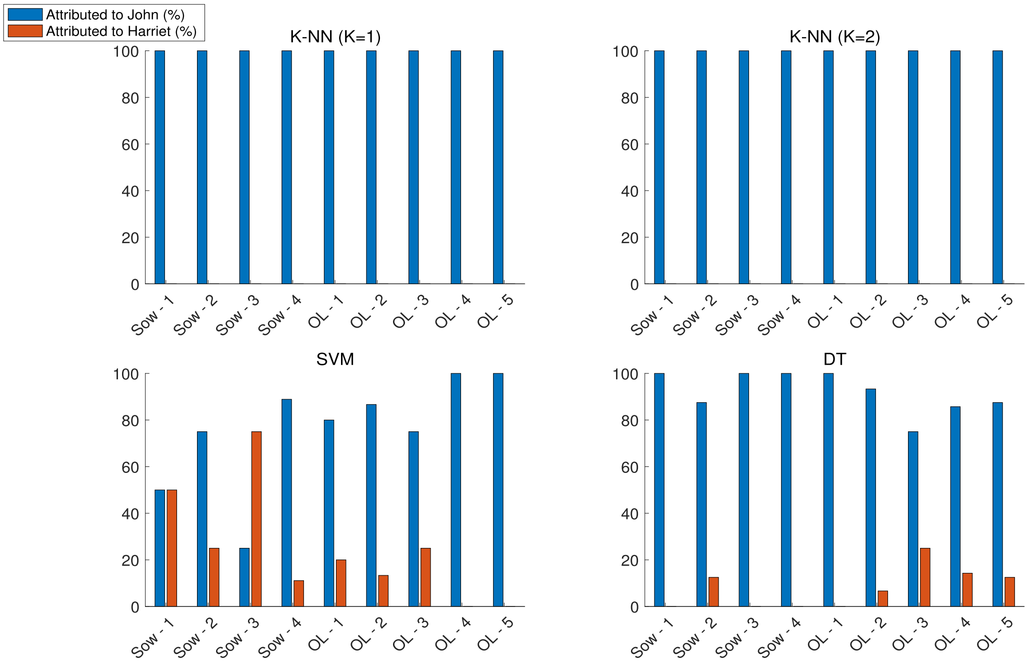


FIGURE 7. Test results. The bar plots show the percentage of attribution into a class by the four models.

of Women. For example, there is no surviving manuscript in either of their hands or one with corrections, notes, suggestions (as is the case with other works by John Stuart Mill). Although our results corroborate the results in [1], suggesting that John Stuart Mill's was the authorial hand, the fact that there are similarities with Harriet Taylor Mill's body of work (without known assistance or help by John Stuart Mill) gives much credence to the claim that these texts were products of earlier collaboration. Most importantly, we need to examine closely the chunks of the test set that are attributed to Harriet

Taylor Mill to bright to the fore, and evaluate the impact of, any possible contribution on her part. However, in this paper we focus on the proposed framework rather than the literary investigation.

VI. CONCLUSION

The main purpose of this work is to create models that are able to learn and distinguish the writing style of two authors. Then we use these models to attribute the authorship of two essays of disputed authorship, bearing strong evidence of collaboration between them.

The feature space we extracted from the training set seems to be adequate to learn the data, since in the validation set we achieve 100% of accuracy. When evaluating the test set with those models, the systems attribute the two essays mostly to John Stuart Mill. Given that Harriet Taylor Mill had died before the publication of the two texts under examination, this result is not surprising. As noted in the Introduction, the disagreement between experts essentially lies between those who accept Harriet Taylor Mill's stylistic influence but reject co-authorship status and those who accept Taylor Mill's co-authorship status despite the lack of corroborating historical evidence. Our results confirm that the texts were written by John Stuart Mill. However, the difficulty in making the attribution makes his claim, that the essays were the product of collaboration to the point of co-authorship, seem more and more credible.

REFERENCES

[1] A. Neocleous and A. Loizides, "Machine learning and feature selection for authorship attribution: The case of mill, Taylor mill and Taylor, in the nineteenth century," *IEEE Access*, vol. 9, pp. 7143–7151, 2021.

[2] D. I. Holmes, "Authorship attribution," *Comput. Hum.*, vol. 28, pp. 87–106, Apr. 1994.

[3] H. Love, *Attributing Authorship: An Introduction*. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[4] E. Stamatatos, "A survey of modern authorship attribution methods," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 60, no. 3, pp. 538–556, Mar. 2009.

[5] T. Neal, K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, and D. Woodard, "Surveying stylometry techniques and applications," *ACM Comput. Surv.*, vol. 50, no. 6, pp. 1–36, Nov. 2018.

[6] J. F. Burrows, "All the way through: Testing for authorship in different frequency strata," *Literary Linguistic Comput.*, vol. 22, no. 1, pp. 27–47, 2007.

[7] M. Eder, "Does size matter? Authorship attribution, small samples, big problem," *Literary Linguistic Comput.*, vol. 30, pp. 167–182, Jun. 2015.

[8] K. Luyckx, *Scalability Issues in Authorship Attribution (Brussels: University Press Antwerp)*. Brussels, Belgium: Univ. Press Antwerp, 2010.

[9] M. Kestemont, K. Luyckx, W. Daelemans, and T. Crombez, "Cross-genre authorship verification using unmasking," *English Stud.*, vol. 93, no. 3, pp. 340–356, May 2012.

[10] A. Severyn and A. Moschitti, "Twitter sentiment analysis with deep convolutional neural networks," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2015, pp. 959–962.

[11] S. Swain, G. Mishra, and C. Sindhu, "Recent approaches on authorship attribution techniques—An overview," in *Proc. Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA)*, Apr. 2017, pp. 557–566.

[12] J. Rudman, "The state of non-traditional authorship attribution studies—2012: Some problems and solutions," *English Stud.*, vol. 93, no. 3, pp. 259–274, May 2012.

[13] J. Rudman, "Stylometrics," in *Cambridge Encyclopedia of the Language Sciences*, P.C. Hogan, Ed. Cambridge, U.K.: Cambridge Univ. Press, 2011, pp. 817–819.

[14] A. Hayward, "John Stuart Mill," *Frazer's Mag.*, vol. 8, no. 48, pp. 663–681, 1873.

[15] H. Cowell, "John Stuart Mill, autobiography," *Blackwood's Edinburgh Mag.*, vol. 115, no. 699, pp. 75–93, 1874.

[16] F. T. Palgrave, "John Stuart Mill's autobiography," *Quart. Rev.*, vol. 136, no. 1, pp. 150–179, 1874.

[17] H. O. Pappe, *John Stuart Mill and the Harriet Taylor Myth*. Melbourne, VIC, Australia: Australian National Univ. Press, 1960.

[18] F. Mineka, "The autobiography and the lady," *Univ. Toronto Quart.*, vol. 32, no. 3, pp. 301–306, 1963.

[19] J. M. Robson, "Harriet Taylor and John Stuart Mill: Artist and scientist," *Queen's Quart.*, vol. 73, no. 2, pp. 167–186, 1966.

[20] J. Stillinger, *Multiple Authorship and the Myth of Solitary Genius*, Oxford, U.K.: Oxford Univ. Press, 1991.

[21] F. A. Hayek, "John Stuart Mill and Harriet Taylor: Their friendship and subsequent marriage," in *The Mill-Taylor Friendship and Related Writings*, S. J. Peart, Ed. Chicago, IL USA: Univ. Chicago Press, 2015, pp. 3–270, 1951.

[22] M. S. J. Packe, *The Life of John Stuart Mill*. London, U.K.: Secker and Warburg, 1954.

[23] G. Himmelfarb, *On Liberty and Liberalism*. New York, NY, USA: A.A. Knopf, 1974.

[24] A. Bain, *John Stuart Mill: A Criticism*. London, U.K.: Longmans, Green and Co, 1882.

[25] H. McCabe, "Harriet Taylor," in *A Companion to Mill*, C. MacLeod and D. Miller, Eds. Hoboken, NJ, USA: Wiley, 2016, pp. 112–125.

[26] A. S. Rossi, "Sentiment and intellect: The story of John Stuart Mill and Harriet Taylor Mill," in *Essays on Sex Equality; John Stuart Mill & Harriet Taylor Mill*, A. S. Rossi, Ed. Chicago, IL, USA: Univ. Chicago Press, 1970.

[27] J. E. Jacobs, "The lot of gifted ladies is hard': A study of Harriet Taylor Mill criticism," *Hypatia*, vol. 9, no. 3, pp. 132–162, 1994.

[28] M. Philips, "The 'beloved and deplored' memory of Harriet Taylor Mill: Rethinking gender and intellectual labor in the canon," *Hypatia*, vol. 33, no. 4, pp. 626–642, 2018.

[29] J. Collins, D. Kaufer, P. Vlachos, B. Butler, and S. Ishizaki, "Detecting collaborations in text comparing the authors' rhetorical language choices in the federalist papers," *Comput. Hum.*, vol. 38, pp. 15–36, Feb. 2004.

[30] E. Dauber, R. Overdorf, and R. Greenstadt, "Stylometric authorship attribution of collaborative documents," in *Cyber Security Cryptography and Machine Learning* (Lecture Notes in Computer Science) vol. 10332, S. Dolev and S. Lodha, Eds. Cham, Switzerland: Springer, 2017.

[31] M. Koppel, J. Schler, S. Argamon, and Y. Winter, "The 'fundamental problem' of authorship attribution," *English Stud.*, vol. 93, no. 3, pp. 284–291, May 2012.

[32] P. Juola, "Authorship attribution," *Found. Trends Inf. Retr.*, vol. 1, no. 3, pp. 233–334, 2006.

[33] J. F. Burrows, *Computation into Criticism*. Oxford, U.K.: Clarendon Press, 1987.

[34] H. Craig, "Contrast and change in the idiolects of Ben Jonson characters," *Comput. Hum.*, vol. 33, no. 3, pp. 221–240, 1999.

[35] L. Campbell, *The Sophistes and Politicus of Plato, With a Revised Text and English Notes* Oxford, U.K.: Clarendon Press, 1867.

[36] W. Lutoslawski, *The Origin and Growth of Plato's Logic, with an Account of Plato's Style and of the Chronology of his Writings*. London, U.K.: Longmans, Green, 1897.

[37] J. T. Temple, "A multivariate synthesis of published platonic stylometric data," *Literary Linguistic Comput.*, vol. 11, no. 2, pp. 67–75, Jun. 1996.

[38] D. L. Hoover, "Multivariate analysis and the study of style variation," *Literary Linguistic Comput.*, vol. 18, no. 4, pp. 341–360, Nov. 2003.

[39] J. S. Mill, *Collected Works of John Stuart Mill*, J.M. Robson, Ed. Toronto, ON, Canada: Univ. Toronto Press, 1963.

[40] H. T. Mill, *Complete Works of Harriet Taylor Mill*. J. E. Jacobs, Ed. Bloomington, IND, USA: Indiana Univ. Press, 1998.

[41] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. theory*, vol. 13, no. 1, pp. 21–27, 1967.

[42] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[43] K. S. Durgesh and B. Lekha, "Data classification using support vector machine," *J. Theor. Appl. Inf. Technol.*, vol. 12, no. 1, pp. 1–7, 2010.

[44] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020.

[45] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, pp. 45–66, Nov. 2001.

[46] L. Wei, B. Wei, and B. Wang, "Text classification using support vector machine with mixture of kernel," *J. Softw. Eng. Appl.*, vol. 5, no. 12, pp. 55–58, 2012.

[47] O. Maimon and L. Rokach, "Decision trees," in *Data Mining and Knowledge Discovery Handbook*. Boston, MA, USA: Springer, 2005, pp. 165–192.

[48] F. Mosteller and D. L. Wallace, *Applied Bayesian and Classical Inference; the Case of the Federalist Papers*, 2nd ed. Berlin, Germany: Springer-Verlag, 1984.

[49] B. Kjell, "Authorship determination using letter pair frequency features with neural network classifiers," *Literary Linguistic Comput.*, vol. 9, no. 2, pp. 119–124, Apr. 1995.

[50] D. I. Holmes, "The federalist revisited: New directions in authorship attribution," *Literary Linguistic Comput.*, vol. 10, no. 2, pp. 111–127, Apr. 1995.

[51] F. J. Tweedie, S. Singh, and D. I. Holmes, "Neural network applications in stylometry: The federalist papers," *Comput. Hum.*, vol. 30, no. 1, pp. 1–10, 1996.

[52] S. Levitan and S. Argamon, "Fixing the federalist: Correcting results and evaluating editions for automated attribution," in *Proc. Digit. Hum., Conf.* Paris, France: Univ. Paris-Sorbonne, 2006, pp. 323–328.

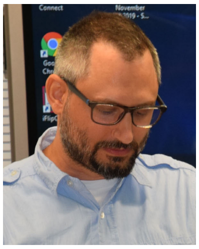
[53] A. Miranda-García and J. Calle-Martín, "The authorship of the disputed federalist papers with an annotated corpus," *English Stud.*, vol. 93, no. 3, pp. *371–390*, 2012.

[54] J. Savoy, "The federalist papers revisited: A collaborative attribution scheme," *Proc. Amer. Soc. Inf. Sci. Technol.*, vol. 50, no. 1, pp. 1–8, 2013.

[55] X. Puig, M. Font, and J. Ginebra, "A unified approach to authorship attribution and verification," *Amer. Statistician*, vol. 70, no. 3, pp. 232–242, Jul. 2016.

ANDREAS NEOCLEOUS was born in Cyprus. He received the bachelor's degree from the University of Pompeu Fabra, Spain, and the Ph.D. degree from the University of Groningen, The Netherlands, in 2016. He studied audio signal processing at the Technical University of Crete, Greece. He has been collaborating with the University of Cyprus (UCY), Cyprus, as a Research Scientist, since 2011, on research programs funded by the EU, the UCY, and the Cyprus Research Promotion Foundation. He is currently a Postdoctoral Researcher at the UCY. He has published articles and presented his work at international conferences. His research interests include digital signal processing, machine learning, and computational intelligence.

GIORGOS KATALIAKOS received the bachelor's (B.A.) degree in political science and history from Panteion University, the master's (M.A.) degree in social and political thought from the University of Sussex, and the Ph.D. degree in theory and philosophy of education from the University of Cyprus with a focus on the issue of curiosity and education in the philosophy of the early modern thinker Thomas Hobbes. His research interests include systematically on epistemology and early modern constitutionalism, education, theory of the state, colonial and postcolonial theory.



ANTIS LOIZIDES received the Ph.D. degree in the history of political thought from the Queen Mary University of London, in 2011. He is currently a Lecturer with the Department of Social and Political Sciences, University of Cyprus. He focuses on utilitarian political thought, with a special interest in John Stuart Mill, James Mill, and their classical influences. He is the author of *James Mill's Utilitarian Logic and Politics* (Routledge, 2019) and *John Stuart Mill's Platonic Heritage: Happiness Through Character* (Lexington Books, 2013). To date, he has published articles in *Utilitas*, *Modern Intellectual History*, *History of Political Thought*, *British Journal for the History of Philosophy*, and *History of European Ideas*. His research interests include political theory, the history of political thought, and the reception of classics.