IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 18, 2020, accepted December 8, 2020, date of publication December 25, 2020,
date of current version January 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3047583

Machine Learning and Feature Selection for Authorship Attribution: The Case of Mill, Taylor Mill and Taylor, in the Nineteenth Century

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

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

The project was supported by the University of Cyprus (start-up internal funding scheme).

ABSTRACT In this article we revisit a dividing issue as regards the corpus of one of the most famous nineteenth-century philosophers: John Stuart Mill. He was the author of two iconic texts in the history of political philosophy: *On Liberty* and *The Subjection of Women*. However, Mill attributed the first to collaboration with Harriet Taylor Mill, his wife, and characterized the second as a work of three minds: his own, his wife's and her daughter, Helen Taylor. Experts disagree on this issue. Most think Mill was too generous sharing authorship credit. We use a training set consisted in manuscripts of the three above mentioned authors, to train a four-class problem (three authors and joint productions). For every manuscript in the training set we extract a set of features that are widely used in text analytics and classification. Then, we apply some pre-processing techniques to normalize the data and to reduce the number of features. Finally, we train three classifiers, namely k-nearest neighbours (k-NNs) with k = 1 and k = 2, support vector machines (SVMs), and decision trees (DTs) to attribute the texts of "disputed" authorship to one of the four potential authors. We routinely run the experiments using different feature sets every time, in order to identify the optimal combination of features that yield the best results on the test set. The best results are achieved with the SVMs, having as input the bigrams features and their principal components. The mean detection rate for all four classes is 100%. Similar results are achieved with the models built with the k-NNs (k = 1) and the DTs. The only classifier that consistently is returning significantly lower results is the k-NN with k = 2. All of the instances in the test set are attributed to John Stuart Mill.

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

I. INTRODUCTION

The need for developing systems that can automatically attribute an author to a given text has a sense of urgency of late, due to the dramatical increment of texts in which their content is somewhat of a public threat and the author is not known – for example, the possible incitement of people to violent behavior, either towards others or one's self, through social media. More broadly, automated Authorship Attribution (AA) of texts has several applications including criminal investigations (e.g. authenticity of suicide notes), identifying the authors of harassing emails and other [1], [2]. Further, AA has received particular attention in the digital humanities. In the history of ideas, authors frequently published their texts anonymously for various reasons: the threat of censorship, prosecution or persecution, to dissasociate a text from one particular individual or even to cheat the reading public. One famous attempt for attributing important eighteenth-century political texts is the work of Mosteller and Wallace on "The Federalist Papers" [3].

In this work, we are investigating three AA questions as regards John Stuart Mill's corpus. John Stuart Mill (1806–1873) was a very famous British philosopher in the nineteenth century. His influence is still visible today in political and social philosophy, the methodology of the social sciences, as well as economic theory.

The first question involves *On Social Freedom*. In 1905, more than thirty years after Mill's death, this essay was found among his papers in his home at Avignon, France. Published two years later under his name, thanks to an attribution by

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

his wife's granddaughter, it is rather unlikely that he was the author. As Rees in [4] convincingly discussed, not only are the ideas untypical of Mill. But also the manuscript's handwriting does not match Mill's. Further evidence strengthen this claim. A draft 1862 letter by Mill was published by Hugh S.R. Elliot in 1910 [5], in which Mill acknowledged receipt of an essay on freedom by an unknown correspondent. Highly regarded at the time, Mill frequently received letters with essays and requests for advice. How could Mary Taylor mistake the author of *On Social Freedom* for Mill? As we shall see, this essay's authorial style is closer to Mill's than either Harriet Taylor Mill's or Helen Taylor's. Although this does not mean that Mill wrote it, it might explain why Mary Taylor mistook it for Mill's in 1905.

The second and third questions involve collaborative work between John Stuart Mill and Harriet Taylor Mill (1807–1858) as well as John Stuart Mill and Helen Taylor (1831–1907). Mill famously shared authorship credit with these two important women in his life for two great works in the history of political philosophy and classical liberalism: *On Liberty* (1859) and *The Subjection of Women* (1869). However, most of Mill's readers, then and now, are not convinced that his acknowledgement is credible. Mill may have exchanged or shared thoughts and ideas with his wife and step-daughter, they argue [6]–[8], but his was ultimately the guiding hand. Our results seem to corroborate this assessment.

There are several ways for building systems for an automated AA. In this work, we consider the problem as a classification task where the training examples consist of texts of known authors.

We use Machine Learning (ML) techniques and we follow a standard procedure for attributing authors to "anonymous" texts: first, we split the dataset into a training set and a test set. In our case we use the "leave-one-out" cross validation method for doing this. Then, we separate every text into chapters and for every chapter in the training set we extract a pre-defined set of features.

It is important to mention that one of the most important features in text analysis are n-grams. However to note a disadvantage, n-grams grow very big, in proportion to the number of training instances. To give an indication on the growth rate of unigrams and bigrams, in Fig. 1 we illustrate the number of bigrams, as they grow when presenting the training instances one after the other, for fold 1. With different colours we indicate the four classes (three authors and joint productions by John Stuart Mill and Harriet Taylor Mill). It can be seen that the growth rate of bigrams in John's and Helen's texts has higher slope than others, meaning that their writing style has richer vocabulary. This observation is also shown in Section III where we provide some statistical analysis of the data. The number of paragraphs and the number of sentences in John's and Helen's texts have bigger variation than others. It is interesting to see in Fig. 1 that two of the essays by Harriet Taylor Mill separate from the rest of her texts. It is interesting because, before running the experiments, we considered excluding these two essays (*Fitzroy's Bill* and *Enfranchisement of Women*) from Harriet Taylor Mill's training corpus due to John Stuart Mill's involvement in the writing process. He took up the role of copy-editor and scriber to his wife (which is also the reason why we did include these essays), but Fig. 1 raises the issue whether his role was more substantial than that. More so when, as we shall see, one of these essays ("Enfranchisement") also appears as a clear outlier (see below in Section III).

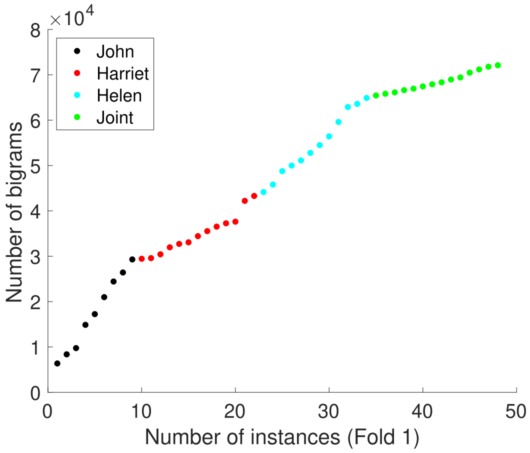


FIGURE 1. Bigram's growth rate while introducing the training instances, one by one.

For the above reason, we introduce another stage in our system to reduce the dimensions of the feature space. Finally in the modelling stage, we train supervised classifiers such as k-NNs and SVMs. The models are then used for attributing the test and the unknown examples into one of the trained classes.

Based on the nature of the problem, we built models using examples of texts from the three above mentioned writers, as well as examples of collaboration (labeled "joint productions") between John Stuart Mill and Harriet Taylor Mill. In this case the models are trained to attribute input data into one of four classes.

Our results suggest that this is a difficult task to solve, which might account for the disagreement between experts, using traditional methods of attribution. Most of the features were not able to achieve satisfying results in learning and classifying validation instances to the correct classes. However, the bigrams together with the SVMs and the DTs returned 100% detection rate for all classes.

In this work we are focusing on feature selection by building models using different combinations of features. We therefore know how the models behave on this level. To the best of our knowledge, this is the first attempt to approach this issue using machine learning and computational text analytics techniques and thus the results of this work can be considered as a state-of-the-art.

The rest of this article is organized as follows: Section II briefly reviews previous research in automated AA; Section III expands on the methods we used to identify authorial patterns as per our research questions; Section IV presents our results. Finally, we offer our conclusions in V and VI.

II. PREVIOUS WORK

A. OVERVIEW

Authorship identification as a scientific task is being studied for more than 130 years now. Mendenhall in 1887 tried to attribute texts in one of the three authors: Bacon, Marlowe, and Shakespeare, using some simple textual features and statistics, [9]. Soon after this work, the community started experimenting adopting more complicated methods to their systems, including feature extraction and techniques that are able to capture the writing style of authors, e.g. [10], [11].

The complexity of the AA task is due to several factors, including the increased number of classes to be modelled and identified, e.g. the number of candidate authors [12]–[15], the different context and genre of texts within a class [16], [17], the big number of features that are now available and need to be handled [19], [25], among others. For example, thousands of features can be extracted from a text, if we consider the frequency of every unique word in a text as a separate feature. This number increases dramatically when we consider the frequencies of tuplets or triplets of words (n-grams). If we consider a classification task with only a few examples characterized by a large number of features we can think that the classifiers will be overfitting the data. Then, if we turn the problem into a multi-class task which is the dominant case in automatic AA, the classifiers will get more difficulties in learning and assigning the data into their respective class.

B. FEATURE SELECTION

Researchers propose several methods for reducing the number of features in an automated AA task. Some apply basic rules on the features space, such as to choose the n most frequent words in the corpus, where n is a parameter to set by the user. For example, Burrows [20] used $n = 100$, Koppel *et al.* [21] used $n = 250$ and Stamatatos [22] used $n = 1000$. The above mentioned method is found to be effective in some scenarios. Another method for dimensionality reduction is to filter out the features that are not discriminative between classes, [23], [24]. This is generally not recommended to be used alone because it might create biased sets into discriminative classes by nature, such as texts that belong to different genres. Other methods include the ‘‘odds ratio’’ [25] and the ‘‘feature instability’’ which essentially are those features that stay relatively unchanged even when the meaning of the text changes, [26]. Further, studies widely use ‘‘principal component analysis’’ for dimensionality reduction, e.g. in [27] and in [20].

C. AUTHORSHIP ATTRIBUTION APPROACHES

One can classify the AA approaches into the way they treat the training set and into the methods that are used to model the writing style of authors and classify unknown texts into one of the learned classes.

In the first case, one way is to concatenate all the training examples of every class (e.g. several texts or manuscripts of the same author) into one instance, [28]. This approach is also called ‘‘profile-based’’ approach. The other way is to treat every example or every paragraph, or chunks of specific number of words in the training set as a separated instance (‘‘instance-based’’ approach), which is the most common case in automated AA, [29], [30].

In the second case, approaches are separated into statistical methods and into machine learning methods. Statistical methods include probabilistic classifiers such as Naïve Bayes or linear discriminant analysis [31], [33]. Machine learning methods include k-NNs, SVMs, artificial neural networks and others, [34], [35].

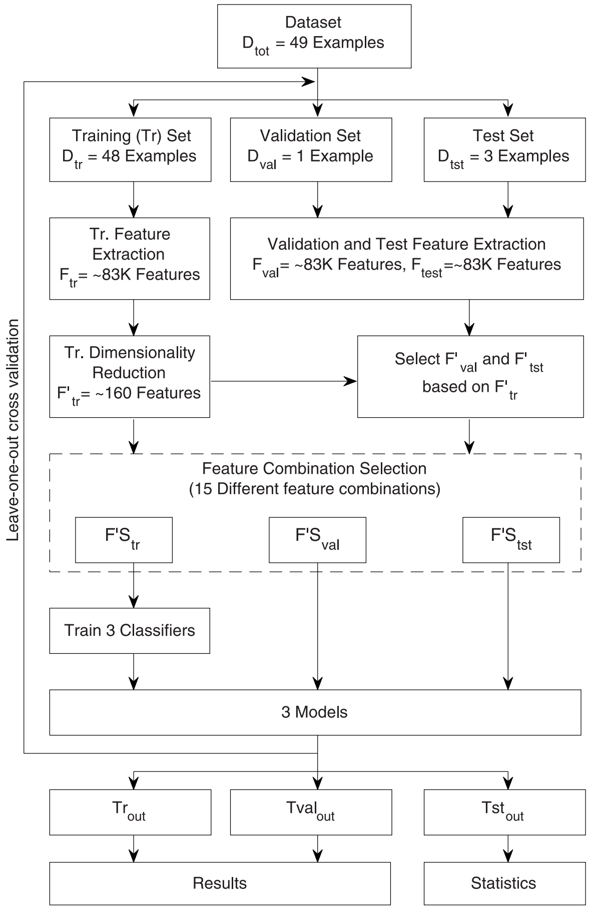


FIGURE 2. The pipeline of our system. It consists of a ‘‘leave-one-out’’ cross validation, feature extraction, dimensionality reduction, training, validation and feature selection.

III. METHODS

A. OVERVIEW

The pipeline of our system is shown in Fig. 2. The initial dataset consists of $D_n = 49$ texts. In this study we apply a ‘‘leave-one-out’’ cross validation approach which means that for every classifier we train, we create D_n models. In each of the above mentioned models, we use one example from the dataset for validation, different every time, and we use the remaining ones for training. The test set consists of the three essays *The Subjection of Women*, *On Liberty* and *On Social Freedom* (See Section I). Therefore, for every fold, we create three sets namely "Training set" (D_{tr}), "Validation set" (D_{val}) and "Test set" (D_{tst}). Then, from every set we extract a number of features (f_{tr}). We apply some filtering to reduce the dimensions of f_{tr} and we use the new filtered feature set (f'_{tr}) as input to train four classifiers. To test these four models, first we extract from the validation and the test sets the same features as done for the F_{tr} set. Then, we select the same features as resulted in f'_{tr}, to create the f'_{val} and f'_{tst} sets. For every fold we compute the detection rates (DR) for both training and test sets and we report the results as an average DR. The results of the test set are reported using statistics. This procedure repeated 15 times using different feature combination each time (e.g. only lexical features or only punctuations, etc).

B. DATA

In Table 1 we present the details of our dataset. We have collected 27 essays by three authors (two by John Stuart Mill that are split in five chapters each, 13 by Harriet Taylor Mill and 12 by Helen Taylor) and 14 essays that are joint productions by John Stuart Mill and Harriet Taylor Mill. These are used to train and validate the system. The test set consists of three essays of unknown authors. In the first two columns of Table 1 we present the author and the titles of the essays used in our dataset. In the third column we provide the year that the essay was written and in the fourth 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.

C. FEATURE EXTRACTION

The features we use for the task at hand are separated in five categories: 1) Counts, 2) Punctuations, 3) CLAWS tags, 4) Normalizations, 5) n-grams (unigrams and bigrams). In the "Counts" category, we extract the following features: 1.1) Number of paragraphs, 1.2) Number of sentences, 1.3) Number of words, 1.4) Average paragraph length, 1.5) Average sentence length, 1.6) Average word length, 1.7) Std paragraph length, 1.8) Std sentence length, 1.9) Std word length, 1.10) Average number of sentences in paragraphs, 1.11) Average number of words in paragraphs, 1.12) Average number of words in sentences, 1.13) Std of number of sentences in paragraphs, 1.14) Std of number of words in paragraphs, 1.15) Std of number of words in sentences, while in the category "Punctuations" we select a list of 17 punctuations. The "CLAWS tags (the Constituent Likelihood Automatic Word-tagging System)" is a list of 138 grammatical tags. This tool is developed by the University Centre for Computer Corpus Research on Language (UCREL) and it is freely available online.[1] In the "Normalizations" category, we apply the following divisions: 4.1) Number of paragraphs/Number of sentences, 4.2) Number of paragraphs/Number of words 4.3) Number of sentences/Number of words, 4.4–4.21 Every punctuation/All punctuations and 4.22–4.161 Every CLAWS tag/All CLAWS tags. The unigrams feature describe the frequency of every word in a text, e.g., how many times a word appears in a document. The bigrams describe the frequency of every consecutive pair of words in a document.

D. DIMENSIONALITY REDUCTION

It is well known that n-grams make for a strong feature in text analytics [1]. If a writer uses the same word more frequently than other writers, then that specific word can be considered a discriminative feature. This can become stronger by examining pairs of consecutive words (bigrams). However, as stated in the Introduction, n-grams are generally very large in size which is proportional to the length of the input text. In our experiments, the average unigram length for all the 49 folds is 11067 features and the average bigram length is 71580 features. It is preferable in machine learning to build models that utilize as less features as possible, for a number of reasons. The most straightforward one is to reduce the computational power needed for training and for feature extraction. Others include complexity in feature extraction, if these require additional effort and equipment.

In this work, we used two methods for dimensionality reduction on CLAWS tags, unigrams and bigrams. The first method is a statistical one and chooses only the features that are highly discriminative between a selected class against the other classes. For example, we identify the features that have non-zero values in more than 50% in the selected class and the sum of the values in that class is greater than $P\%$ of the sum of the whole feature, having $P = 70$ for CLAWS tags, $P = 90$ for unigrams and $P = 92$ for bigrams. The above mentioned percentages are chosen heuristically. In Fig. 3 we illustrate two examples on how the filtered data of the first method are illustrated for two pairs of selected features. In the upper subplot, we plot the word "obedience" against the word "benefits". These two words appear in John Stuart Mill's texts more frequently than in the texts of the other authors. Similarly, in the lower sublpot, the words "commited" and "jury" separate the joint productions texts.

The second method is to choose the first principal components that sum up together up to 95% of all the components. To understand better how these data are distributed, in Fig. 4, we present six scatter plots. The first plot shows the distribution of the number of paragraphs against the number of sentences for all the classes. This pair of features is not strong because it contains raw values, meaning that it lacks of any normalizations and therefore they are dependent of the length of the input texts. However, we choose to illustrate them to get an idea how the raw data are distributed. Some first observations are that Harriet Taylor Mill's texts consist in short numbers of paragraphs and numbers of sentences while John Stuart Mill's texts appear to have bigger deviations with

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

TABLE 1. The dataset used for training and test.

Author	Essay name	Year	Number of words	Training or Test
John Stuart Mill	Utiliarism (Ch. 1)	1861	1857	Training
John Stuart Mill	Utiliarism (Ch. 2)	1861	8552	Training
John Stuart Mill	Utiliarism (Ch. 3)	1861	3497	Training
John Stuart Mill	Utiliarism (Ch. 4)	1861	2898	Training
John Stuart Mill	Utiliarism (Ch. 5)	1861	9796	Training
John Stuart Mill	Considerations on Representative Government (Ch. 1)	1861	4374	Training
John Stuart Mill	Considerations on Representative Government (Ch. 2)	1861	7586	Training
John Stuart Mill	Considerations on Representative Government (Ch. 3)	1861	6923	Training
John Stuart Mill	Considerations on Representative Government (Ch. 4)	1861	4124	Training
John Stuart Mill	Considerations on Representative Government (Ch. 5)	1861	6091	Training
Harriet Taylor Mill	Australia	1831	241	Training
Harriet Taylor Mill	German Prince	1832	225	Training
Harriet Taylor Mill	Manners	1832	1331	Training
Harriet Taylor Mill	Hampden	1832	2635	Training
Harriet Taylor Mill	Mirabeau	1832	1416	Training
Harriet Taylor Mill	Plato	1832	623	Training
Harriet Taylor Mill	French Revolution	1832	2536	Training
Harriet Taylor Mill	Seasons	1832	1613	Training
Harriet Taylor Mill	Conformity	MS c1831	1934	Training
Harriet Taylor Mill	Laconicisms	MS c1832	1652	Training
Harriet Taylor Mill	Alroy	MS c1833	601	Training
Harriet Taylor Mill	The Enfranchisement of Women	1851	10012	Training
Harriet Taylor Mill	Fitzroy’s Bill	1853	2372	Training
Helen Taylor	The Education of Women	c1860	1976	Training
Helen Taylor	Women and Criticism	1866	3728	Training
Helen Taylor	The Ladies’ Petition	1867	7590	Training
Helen Taylor	Women’s Rights as Preached by Women	1881	2727	Training
Helen Taylor	On Fox-Hunting	1870	2495	Training
Helen Taylor	On T. More	1870	3519	Training
Helen Taylor	Paris and France	1871	3717	Training
Helen Taylor	New atttack on Toleration	1871	4555	Training
Helen Taylor	Greece and the Greeks	1863	6532	Training
Helen Taylor	Personal Representation	1865	9345	Training
Helen Taylor	Too Late and Too Soon	1873	1695	Training
Helen Taylor	T.H. Buckle Biographical Note	1872, 1. ix-xvii	2920	Training
Joint Production	The Suicide of Sarah Brown	1846	1342	Training
Joint Production	The Case of Susan Moir	1850	733	Training
Joint Production	Questionable Charity	1850	792	Training
Joint Production	Wife Murder	1851	1250	Training
Joint Production	The Acquittal of Captain Johnstone	1846	855	Training
Joint Production	Dr. Ellis”s Conviction	1846	1138	Training
Joint Production	The Case of Private Matthewson	1846	1241	Training
Joint Production	The Case of William Burn	1846	1120	Training
Joint Production	The Case of the North Family	1846	1631	Training
Joint Production	Corporal Punishment	1849	1144	Training
Joint Production	The Case of Mary Ann Parsons 1-2	1850	2457	Training
Joint Production	The Case of Anne Bird	1850	1774	Training
Joint Production	The Law of Assault	1850	1695	Training
Joint Production	Punishment of Children	1850	866	Training
Unknown	The Subjection of Women (Ch. 1)	1869	12169	Test
Unknown	The Subjection of Women (Ch. 2)	1869	8868	Test
Unknown	The Subjection of Women (Ch. 3)	1869	12717	Test
Unknown	The Subjection of Women (Ch. 4)	1869	9963	Test
Unknown	On Liberty (Ch. 1)	1859	5607	Test
Unknown	On Liberty (Ch. 2)	1859	16024	Test
Unknown	On Liberty (Ch. 3)	1859	7830	Test
Unknown	On Liberty (Ch. 4)	1859	7881	Test
Unknown	On Liberty (Ch. 5)	1859	9354	Test
Unknown	On Social Freedom (Ch. 1)	1907	965	Test
Unknown	On Social Freedom (Ch. 2)	1907	992	Test
Unknown	On Social Freedom (Ch. 3)	1907	2344	Test
Unknown	On Social Freedom (Ch. 4)	1907	4571	Test

respect to these two features. Also, it is shown that their joint productions distribute closer to the cluster of Harriet Taylor Mill while Helen Taylor appears to be closer to the cluster of John Stuart Mill. There is a clear outlier of one of the texts of Harriet Taylor Mill. The essay *The Enfranchisement of Women* (1851) was attributed 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, amanuencis and copy-editor to his wife in the process

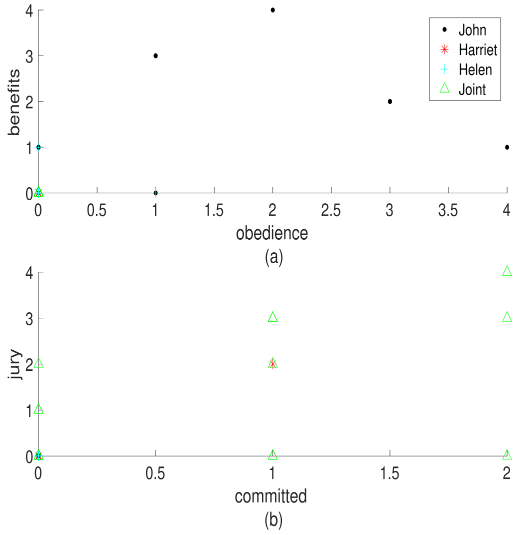


FIGURE 3. Scatter plots of pairs of words that are highly discriminative for (a) John Stuart and (b) for joint productions.

of writing. Some contemporary critics, however, thought that this essay was a poor immitation, a parody, of John Stuart Mill's style [32].

In Fig. 4b we illustrate a scatter plot of the same features mentioned above, for John Mill's, Harriet Taylor Mill's and the test texts. In this case Test 1 and Test 3 are closer to the cluster of John Mill while Test 2 is closer to the cluster of Harriet Taylor Mill.

In Fig. 4c we show the distribution of the first two unigrams PCA components. Here the distance between John Mill and Harriet Taylor Mill is bigger but nevertheless, the observations are consistent with the ones in Fig. 4a. We can still see the outlier in Harriet Taylor Mill's data and this strengthens the argument that the specific text might be influenced or co-authored by John Mill. In 4c, we can still see that Test 2 is closer to the cluster of Harriet Taylor Mill. This is also consistent with our results, presented in Section IV-D.

Figs. 4e and 4f illustrate the first two bigrams PCA components. n-grams are highly non-linear features and therefore the PCA of those features is still non-linear. This non-linearity can be partially illustrated. Even though the classes are separated nicely, we can see that two texts of Harriet Taylor Mill are distributed in a separate cluster, while one of the texts by John Mill distributes closer in Harriet Taylor Mill's cluster. Similarly, three texts of Helen Taylor are distributed in joint productions cluster, which they both interestingly separate from the rest of the data.

Table 2 shows the sizes of the features before and after the dimensionality reduction for both methods 1 and 2.

E. TRAINING SETS AND MODELLING

1) TRAINING SETS

In this work we focus on testing several feature sets to identify the optimal set that can better distinguish the four requested classes. In the first row of Table 3 we present the names of the feature sets. It includes the entire feature set, the normalized, the filtered, and the n-grams both separately and combinations of them.

TABLE 2. This table shows the size of the features before and after two different methods are applied for dimensionality reduction.

	CLAWS	Unigrams	Bigrams	All
Initial feature size	139	11132	72114	83576
Statistical method	1	7	824	1023
PCA method	1	5	27	224

TABLE 3. The different feature groups and their sizes that are tested as inputs to the three classifiers.

Feature Set Name	Feature set size
All features	83576
Counts	15
Punctuations	18
CLAWS tags	139
Normalizations	158
Unigrams	11132
Bigrams	72114
CLAWS filtered - Statistical	1
Unigrams filtered - Statistical	7
Bigrams filtered - Statistical	824
CLAWS filtered - PCA	1
Unigrams filtered - PCA	5
Bigrams filtered - PCA	27
All filtered	865
Normalizations and All filtered	1023

2) MODELLING

We choose three widely used classifiers to approach this problem namely a) k-nearest neighbours (k-nn) with $k = 1$ and $k = 2$, support vector machines (SVMs) and decision trees. All of the above methods use as input vectors representing instances, where every element of that vector represents one feature. Every instance is annotated with a label representing the class that belongs in.

a: K-NEAREST NEIGHBOURS (K-NNs)

K-nearest neighbours is one of the most commonly used classifiers and it is conceptually simple. One of the first articles introducing the "nearest neighbour rule" is the one of Cover and Hart [36] and it belongs to the non-parametric techniques. An unknown instance (P) is assigned to the class of its closest data point (k) in the training set. This is done by calculating all the distances between instance P and all the instances in the training set. For $k > 1$, the instance P is assigned to the class of the majority of the closest data points. The most common similarity measures are the "Euclidean distance", the "Manhattan distance" and the "Hamming distance".

One of the drawbacks of the k-NNs is that it can become computationally expensive and thus time consuming to compute all the distances between the data points and a test case, when it comes to big datasets. Also, if the data are not standardized to the same or similar scale, it might affect the importance of every feature, in a wrong way. For example, if calculating distances on features with big numbers, then these distances will probably be bigger than the

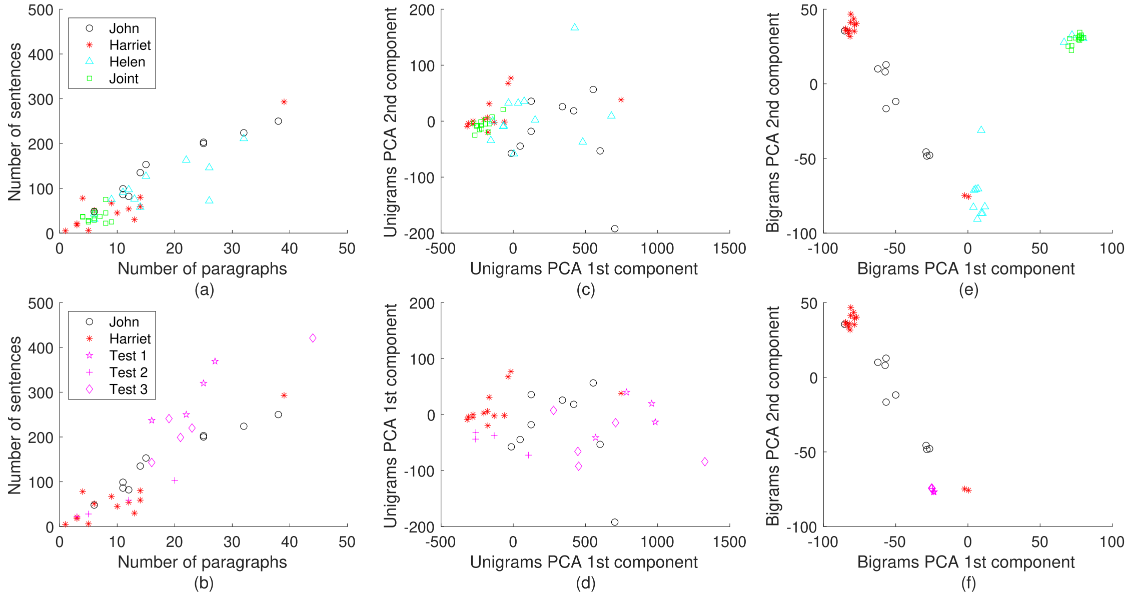


FIGURE 4. **Visualization of the training and the test data. The abbreviations Test 1, Test 2 and Test 3 on the legend stand for the essays: *The Subjection of Women, On Social Freedom* and *On Liberty*, respectively.**

distances of features with smaller numbers. For this reason, the input space of such distance-based methods is better to be normalized.

b: SUPPORT VECTOR MACHINES (SVMs)

The support vector machines are considered by the community as one of the most powerful classifiers. The input data are presented to the algorithm which tries to identify a hyperplane or hyperplanes that separate better the data of different classes. In its simple form, this hyperplane is a linear function. Non-linear functions can be used as well to discriminate better data that do not follow normal distributions [37]. This however, can result into overfitting if the kernel of the function becomes relatively big. New examples are classified based on their largest distance to the hyperplane. A detailed explanation of the SVM are found in [38] and [39], while examples with applications for text classification using SVMs are found in [40] and [41].

c: DECISION TREES (DT)

Decision trees can be used for classification, or for regression problems. They consist of a root node which is the start of the tree, a number of branches and a number of leaf nodes.

The first step in the training phase is to identify the feature that has the highest separability between the defined classes and assign this feature in the root node. This is frequently done by calculating the ‘‘Gini impurity formula’’, but it can be also done by considering some discriminative probabilities or statistics between the classes such as resultant entropy reduction or information gain. As soon as the first feature is found, it is then used to split the data into two categories, based on the feature that is defined and assigned in the root node. This splitting is done by setting a binary decision rule (yes or no) for categorical features or using the mean squared error (MSE) of the feature for continuous variables. In the latter case, the splitting criterion is chosen to be the one which yields the lowest MSE and it carries the concept of the branch in the tree.

The above procedure then applies to lower levels of the tree by splitting the data into sub-categories using the remaining features hierarchically, based on the level of their separability. A node becomes a leaf node when the data cannot be split into further sub-categories, e.g. there are no more features, or all the data are assigned in one sub-category.

A test instance is ‘‘passed’’ through the tree and a decision is made accordingly by meeting all the rules that are defined in the branches.

IV. RESULTS

A. OVERVIEW

In this article we have selected essays by three different writers of the nineteenth century to use them as input for training three classifiers. Because this dataset is relatively small, we decided to apply a ‘‘leave-one-out’’ cross validation procedure. We thus present the results as averages of all the 49 folds. Furthermore, we applied a feature test by feeding the classifiers with different groups of features, as explained in Section III-E.

B. TRAINING SET

The majority of the classifiers used in this study are able to learn the training set and they return 100% detection rates (DRs) for all the four classes. More specifically, the k-NNs appear to be the most suitable to learn the data for most of the feature sets, while for $k = 2$, the training set is not learned well. The SVMs and the DTs work better when they are trained with the n-grams and their PCA features.

C. VALIDATION SET

The best average results of all the 49 folds for the test set are achieved with the SVMs using the bigrams and their PCA, returning a 100% DRs for all the classes, following the DTs with using the PCA of the bigrams. It is interesting to mention that k-NNs in the training set return very good results, but fail to predict test instances correctly. This is probably due to overfitting the data. In Table 4 we present the mean DRs of the SVMs for all the feature groups tested.

TABLE 4. The mean DRs of the SVMs for all the feature groups tested.

Feature Set Name	Harriet	Helen	John	Joint
All features	15	25	80	0
Counts	8	42	40	7
Punctuations	8	17	90	0
CLAWS tags	8	17	60	0
Normalizations	0	0	0	0
Unigrams	0	0	90	0
Bigrams	100	100	100	100
CLAWS filtered - Statistical	8	0	0	0
Unigrams filtered - Statistical	0	17	60	0
Bigrams filtered - Statistical	100	83	80	100
CLAWS filtered - PCA	31	8	40	7
Unigrams filtered - PCA	8	0	60	7
Bigrams filtered - PCA	100	100	100	100
All filtered	54	33	90	7
Normalizations and All filtered	46	33	90	7

D. TEST SET

By looking at the results on the three essays that are used as test set, the vast majority of the methods, for all the folds and for all the feature sets tested, attribute the authorship to John Stuart Mill.

V. DISCUSSION

In this article we try to build models that can separate unknown texts into one of three possible authors or in a class called ‘‘joint productions’’.

We consider this four-class classification task to be somewhat complicated to solve because it involves three writers that had close relationship to each other and therefore they were frequently sharing their thoughts. This creates evidence that there was an influence in their writings too and therefore possible correlations in the dataset.

The number of features become very big and this number is proportional to the number of input instances. It is important in such cases to apply dimensionality reduction techniques to allow the system work faster and more efficient. From our results we found out that the bigrams and their PCAs return 100% detection rate for all the classes and therefore the 27 selected PCAs (out of 72K bigrams) are enough to capture all the information needed to train a classifier.

Validating our system with the test examples in our dataset, we observed that all of the three essays are attributed to John Stuart Mill. This was the expected result given that *On Liberty* and *The Subjection of Women* were published with his, but not Harriet Taylor Mill’s, finishing touches. Not only had Helen Taylor no acknowledged role in *On Liberty*; but also her role, contrary to John Mill’s claims, must have been mininal in *The Subjection of Women*. This is something which no scholar has doubted to date. Finally, given Mill’s status in Victorian England, and the misattribution by Harriet Taylor Mill’s granddaughter, it is no wonder that *On Social Freedom* is classed along with other works by Mill rather than Harriet Taylor Mill or Helen Taylor.

In a future work, we aim to test our method by reducing the number of classes (e.g. one against all) and to increase the number of examples by taking each paragraph as a separate example. Also, we will apply unsupervised techniques to observe any meaningful clusters in the data.

VI. CONCLUSION

The main purpose of this work is to identify the best feature combination for learning and classifying texts from three different authors. From a large variety of combinations of features that we test, we identify the bigrams as the most suitable. The SVMs and the DTs seem to be the most powerful classifiers to solve this particular task. Future work will include better analysis on the training set, introducing one instance for every sentence. Also, simpler models with two classes may return more robust results.

REFERENCES

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

[2] I. N. Bozkurt, Ö. Bağlioğlu, and E. Uyar, “Authorship attribution: Performance of various features and classification methods,” in *Proc. 22nd Int. Symp. Comput. Inf. Sci. (ISCIS)*, 2007, pp. 158–162.

[3] F. Mosteller and L. W. David, “Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed *Federalist* Papers,” *J. Amer. Stat. Assoc.*, vol. 58, no. 302, pp. 275–309, 1963.

[4] J. C. J. Rees, *Stuart Mill’s on Liberty*, G. L. Williams, Ed. Oxford, U.K.: Oxford Univ. Press, 1985.

[5] H. S. R. Elliot, Ed., *The Letters of John Stuart Mill*, vol. 2. London, U.K.: Longmans, 1910.

[6] 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, Aug. 1994.

[7] J. M. Robson, “Harriet Taylor and John stuart mill: Artist and scientist,” *Queen’s Quart.*, vol. 73, no. 2, pp. 167–186, 1966.

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

[9] A. Morgan, “The characteristic curves of composition,” *Science*, vol. 13, no. 313, p. 92, Feb. 1889.

[10] D. Holmes, “Authorship attribution,” *Comput. Humanities*, vol. 28, no. 2, pp. 87–106, 1994.

[11] J. Rudman, “The state of authorship attribution studies: Some problems and solutions,” *Comput. Humanities*, vol. 31, pp. 351–365, Jul. 1998.

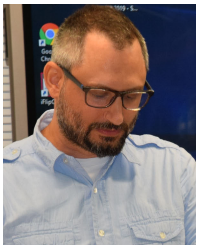
[12] H. van Halteren, “Linguistic profiling for author recognition and verification,” in *Proc. 42nd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2004, pp. 1–9.

[13] S. Argamon, M. Šarić, and S. S. Stein, “Style mining of electronic messages for multiple authorship discrimination: First results,” in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (ACM SIGKDD)*, 2003, pp. 475–480.

[14] M. Koppel, J. Schler, S. Argamon, and E. Messeri, “Authorship attribution with thousands of candidate authors,” in *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2006, pp. 659–660.

[15] K. Luyckx and W. Daelemans, "Authorship attribution and verification with many authors and limited data," in *Proc. 22nd Int. Conf. Comput. Linguistics (COLING)*, 2008, pp. 513–520.
[16] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Text genre detection using common word frequencies," in *Proc. 18th Conf. Comput. Linguistics*, 2000, pp. 1–7.
[17] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Automatic text categorization in terms of genre and author," *Comput. Linguistics*, vol. 26, no. 4, pp. 471–495, Dec. 2000.
[18] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, Mar. 2003.
[19] J. Houvardas and E. Stamatatos, "N-gram feature selection for authorship identification," in *Proc. Int. Conf. Artif. Intell., Methodol., Syst., Appl.* Berlin, Germany: Springer, 2006, pp. 77–86.
[20] J. F. Burrows, "Not unles you ask nicely: The interpretative nexus between analysis and information," *Literary Linguistic Comput.*, vol. 7, no. 2, pp. 91–109, 1992.
[21] M. Koppel, J. Schler, and E. Bonchek-Dokow, "Measuring differentiability: Unmasking pseudonymous authors," *J. Mach. Learn. Res.*, vol. 8, pp. 1261–1276, Jun. 2007.
[22] E. Stamatatos, "Authorship attribution based on feature set subspacing ensembles," *Int. J. Artif. Intell. Tools*, vol. 15, no. 5, pp. 823–838, Oct. 2006.
[23] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, nos. 1–2, pp. 273–324, Dec. 1997.
[24] J. Li, R. Zheng, and H. Chen, "From fingerprint to writeprint," *Commun. ACM*, vol. 49, no. 4, pp. 76–82, Apr. 2006.
[25] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, Mar. 2003.
[26] M. Koppel, N. Akiva, and I. Dagan, "Feature instability as a criterion for selecting potential style markers," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 57, no. 11, pp. 1519–1525, 2006.
[27] J. N. G. Binongo, "Who wrote the 15th book of oz? An application of multivariate analysis to authorship attribution," *Chance*, vol. 16, no. 2, pp. 9–17, Mar. 2003.
[28] F. Mosteller and D. L. Wallace, *Inference and Disputed Authorship: The Federalist*. Reading, MA, USA: Addison-Wesley, 1964.
[29] C. Sanderson and S. Guenter, "Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Morristown, NJ, USA, 2006, pp. 482–491.
[30] G. Hirst and O. Feiguina, "Bigrams of syntactic labels for authorship discrimination of short texts," *Literary Linguistic Comput.*, vol. 22, no. 4, pp. 405–417, Sep. 2007.
[31] R. Clement, "Ngram and Bayesian classification of documents for topic and authorship," *Literary Linguistic Comput.*, vol. 18, no. 4, pp. 423–447, Nov. 2003.
[32] F. Palgrave and T. John Stuart, "Mill's autobiography," *Quart. Rev.*, vol. 136, no. 1, pp. 150–179, 1874.
[33] D. Madigan, A. Genkin, D. Lewis, S. Argamon, D. Fradkin, and L. Ye, "Author identification on the large scale," in *Proc. CSNA*, 2005, pp. 1–20.
[34] J. Diederich, J. Kindermann, E. Leopold, and G. Paass, "Authorship attribution with support vector machines," *Appl. Intell.*, vol. 19, nos. 1–2, pp. 109–123, 2003.
[35] G. Teng, M. Lai, J. Ma, and Y. Li, "E-mail authorship mining based on SVM for computer forensic," in *Proc. Int. Conf. Mach. Learn. Cybern.*, vol. 2. Washington, DC, USA, Aug. 2004, pp. 1204–1207.
[36] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.
[37] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[38] 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.
[39] 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.
[40] 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.
[41] 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.

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



ANTIS LOIZIDES received the Ph.D. degree in history of political thought from Queen Mary University of London, in 2011. He is currently a Lecturer with the Department of Social and Political Sciences, University of Cyprus. He has authored *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 authored or coauthored articles in *Utilitas, Modern Intellectual History, History of Political Thought*, the *British Journal for the History of Philosophy*, and *History of European Ideas*. He focuses on utilitarian political thought, with a special interest in John Stuart Mill, James Mill, and their classical influences. His research interests include political theory, the history of political thought, and the reception of classics.