ICS 3201 Knowledge Discovery and Management Assignment 2015 Author Identification

Nigel Grech
University of Malta
nigelgrech.13@um.edu.mt

ABSTRACT

The ability to automatically detect the author of a particular text can be highly useful in this day and age were individuals can publish documents with complete anonymity on the Internet. This paper describes the work carried out in the development of an authorship classifier as a means of experimenting with possible feature sets and machine learning models. The paper also delves into the background of this field examining and comparing work which has already been carried out. The work carried out for this project focuses on the use of lexical and word based feature sets which have been used to train several different types of models and the evaluation of these models, as well as analysis and discussion of the results obtained.

1. INTRODUCTION

Authorship identification or authorship categorization is the task of identifying the author of a particular text by means of using exemplar texts [1]. Author identification is not a new field and has been around since as early as 1887 when work was carried out on the plays written by Shakespeare, however this early work was not an automated process rather one done manually. It wasn't until the 1960's that this work moved from human experts to machines.

The invention of the World Wide Web has provides opportunities for authors to publish anything under any name, researchers in the field of author identification have already discussed how some individuals may take advantage of this anonymity and use it for their own gain, be it leaking of official document, the spread of hate speech or to commit acts of fraud [1] [2]. Given the potential abuse of publishing with anonymity systems with the ability to perform author classification to a high degree of accuracy would provide a means of bringing such individuals to justice. However such systems also has the potential for abuse, and may be used to identify individuals who have published documents anonymously yet have not committed any crimes but were simply exercising their right to free speech.

The second section of this paper provides a brief history of this field as well as a description of current proposed solutions which tackle the following two major issues. The first being that of feature selection, what features within a document are valuable. It is important to note that the majority of these features should be context free allowing the classifiers to correctly identify an author even if they write about racially different subject matters. The next problem is that of the best way to learn the features of the data provided which models provide the best results, and which are capable of providing good results with large scale data.

In section three we discuss the features chosen for the experiments and how they were extracted from the data set along with the machine learning algorithms used to train authorship classifiers. Section three also discusses the how the classifiers were evaluated.

The results obtained from the evaluation of the classifiers generated are shown in section four along with a discussion and interpretation of the results. In the final section we review possible future work and improvements which can be made to improve the classifiers built.

2. BACKGROUND

This section is divided into three subsections the first is a brief history of authorship identification and its application, the second discusses feature selection with an emphasis on stylometry. The final section reviews the models and techniques which have been used in related work to generate authorship classifiers.

2.1 A Brief History

Authorship classification has been around since the 19th century starting with the work published by Medenhall in 1887 which focused on the works of Shakespeare [3] and the work of Mascol on the gospels of the New Testament. The main idea behind their work being that authors can be distinguished based on the distribution of sentences and word lengths [4].

In the 1930s Yule and Zipf applied statistical techniques to the problem [3], Yule having developed the Yule's K measure to provide a statistical measure of vocabulary richness and Zipf having found the Zipf's distribution and Zipf's law [4]

It wasn't until the 1960s when a seminal piece of work by Mosteller and Wallace introduced a method which used a Bayesian model to analyse function words. This work evaluated the authorship of the famous "Federalist Papers" [3]. This work sparked a line of research that continued well into the late 1990s, referred to as stylometry.

Since the 1990s century the volume of available online texts has grown exponentially and continues to grow. This growth of available resources resulted in the development of efficient information retrieval, machine learning and natural language processing techniques, algorithms and tools [3]. Furthermore this has resulted in the realization of the potential applications for this line of research, such as intelligence, law and forensics [3].

2.2 Features for Authorship Identification

One of the major challenges in authorship identification is that of compiling a set of features which describe an author's writing style in a context free manner. This is a crucial component of an authorship classifier and can mean the difference between an accurate classifier and a bad one.

Stylometry, the statistical analysis of writing style [5], is generally used for this line of research since it provides the necessary data required while remaining context free. Zheng et al. [6] provide a taxonomy for authorship analysis, which has been divided into four sections; Lexical, Syntactical, Structural and Content-specific. The following is a description of these categories based on the taxonomy provided by Zheng et al. [6].

Lexical features, within this category a further two categories may be found character and word based lexical features, where the former are statistics calculated based on the characters found in a text such as the total number of characters, frequency of letters and total number of occurrences of each digit. The word based features are statistical measures based on the words found in a text such as total number of words, total number of unique words and average sentence length [6].

Syntactical features, as the name implies they describe the syntax of the text including such features as function words, frequency of punctuation and frequencies of parts of speech [6].

Structural features describe the way an author has structured the text such as number of lines, number of sentences and number of paragraphs. Note that based on the data set being used to train a classifier it may be impossible to use such features since this information may not be available [6]. Structural features may also comprised of technical features such as file extensions and styling of the text these kinds of features can be rather valuable for online texts [5].

Content-specific features while it is important to use context free features the addition of a limited set of content specific features may help to identify the author. It is important to note however that the set of words to be used in such a case will be dependent on the particular case in question and may be radically different across different domains [6].

2.3 Classification Techniques

Classification using the stylometric features described above may be used to train two types of classifiers Supervised and Unsupervised. The first work with labelled data, author label and feature vector, while the second used unlabelled data [5].

Supervised techniques include the following;

- Neural Networks (NN)
- K Nearest Neighbours (KNN)
- Decision Trees
- Support Vector Machines (SVM)

Unsupervised techniques are mainly comprised principal component analysis (PCA) and cluster analysis, these techniques prove to work well in problems that involve large feature sets [5].

3. METHODOLOGY

This section describes the choices made in designing the experiments, some implementation details of setting up the experiments and the methods of evaluation chosen.

The data which was used for these experiments is a subset of RVC1, a corpus of newswire stories made available by Reuters ltd and can be accessed through [7]. The documents within this data set lack structural features and were selected in such a way as to minimize the topic factor. The implications of these two features of the data set are discussed below in the section pertaining to the features chosen.

3.1 Experiment Design

The task of this project was to experiment with different feature sets and classifiers. With respect to feature sets it was chosen to work with Lexical Features with the aim to see the effect

Feature sets	Label
Character Based	F1
Word Based	F2
F1 + F2	F3

Table 1 Feature Sets

on the accuracy of the classifiers generated using three combinations of lexical features. These sets are listed in table 1. F1 consists of character based lexical features, table 3 contains a full list of the features in this set. F2 consists word based features and a full list of the features used can be seen in table 2. In choosing these set we were inspired by the work done by Zheng et al. [6] however it is important to note that we within the F1 set we also included frequency of punctuation and total number of sentences where as in [6] the former was classified as syntactical and the latter structural. It should also be noted that frequency of function words was also included within the F2 set whereas in [6] it was classified as syntactical. The features described in the sets would then be extracted from the text and compiled as vectors of real numbers, with one vector representing one document.

Due to the fact that the topic factor within the data set was minimal it was chosen not to include context specific features. another important not is that after an examination of some of the data there seemed to be no structure to the files which was intended by the authors, therefore the assumption was taken that the structure of the files was not an intentional (by the author) but rather a side-effect of the data collection process. Hence it was chosen to not include structural features.

The next decision in regards to the experiments was that of which classifiers to train the following classifiers were chosen; KNN, with K being 5, 10 and 15, SVC (multiclass SVM, with multiclass support handled with the one-vs-one schema), linear

SVM (same as SVC however with a linear kernel) and a Gaussian Naive Bayes (GNB) (Naïve Bayes algorithm where the likelihood of the features is assumed to be Gaussian) [8].

Word Based Features
Total number of words (M)
Total number of short words / M
Total number of characters in words
Average word length
Average sentence length in terms of character
Average sentence length in terms of word
Total different words / M
Yule's K measure
Word length frequency distribution /M
Frequency of function words

Table 2 - Word based features

Character Based Features
Total number of characters (C)
Total number of alphabetic characters /C
Total number of upper-case characters /C
Total number of digit characters /C
Total number of white-space characters /C
Frequency of letters
Frequency of special characters
Frequency of punctuations
Total number of sentences

Table 3 - Character based features

Each classifier was trained with each one of the feature sets for varying numbers of authors, this was done to examine how the models will behave as they scale. The number of authors used were 5, 25 and 50. Resulting in a total of 54 models being trained and tested. These models were then evaluated by calculating their precision, recall and F score.

The main aim of these experiments is to observe the following:

1. The effect of the number of authors on the classifiers
2. Which of the feature set classifier combinations performs the best.

It should be noted that the goal is not to achieve high levels of performance form the classifiers but rather to observe trends in the performance based on the feature sets used and the number of authors trained.

3.2 Implementation Details

It was chosen to implement this project using Python and the SciKit-Learn library [8], which is a library that provides machine learning in python.

A function was written to first extract all the features from the texts, then these values were set as vectors and were scaled using an inbuilt function in SciKit (after some experimentation it was found to give better results).

Evaluation was also carried out using a function provided by SciKit.

It is important to note that the goal of these experiments is not to achieve a high level of accuracy in the classifiers but rather to observe the trends of the classifiers as the number of authors increases and as the feature sets change.

4. RESULTS

After running the experiments the results in tables 4, 5 and 6 were recorded for 5, 25 and 50 authors respectively. The first thing which become apparent is that as the number of authors increases the precision, recall and f-score of the classifiers gets very low as can be seen in figure 1. Also note that this trend is common for all the feature sets indicating that lexical features on their own are not a rich enough feature set to perform classification over a large number of authors.

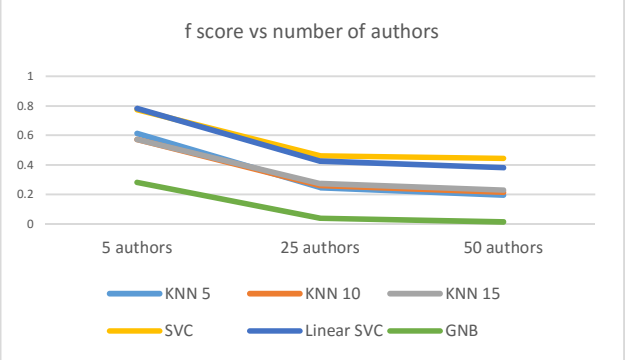

Figure 1 F3 fs-scores vs number of authors

The two classifiers which performed the best are the SVC and Linear SVC classifiers even though both of them degraded with increasing number of authors. An interesting observation is that the linear SVC degraded less than the SVC and with a larger amount of authors outperformed the SVC and vice versa with a smaller number of authors. As can be seen in figure 2. This hints that any future work should focus mainly on the use of SVMs and SVCs to find the right configuration to produce the best results.

The feature set which performed the best was F1 which was counter intuitive since F3 was expected to perform the best since it had the richest feature set. This said the SVC always out performed every other classifier given the F3 feature set, as can be seen in figure 3. This is a curious anomaly and may hit at the fact that the data might be very similar.

Number of authors	Precision			Recall			F Score		
Feature sets	F1	F2	F3	F1	F2	F3	F1	F2	F3
KNN-5	0.686	0.561	0.671	0.664	0.508	0.623	0.660	0.487	0.614
KNN-10	0.675	0.596	0.662	0.668	0.508	0.583	0.662	0.489	0.572
KNN-15	0.77	0.613	0.678	0.647	0.532	0.583	0.640	0.512	0.572
SVC	0.75	0.749	0.794	0.755	0.719	0.772	0.760	0.724	0.773
Linear SVC	0.203	0.710	0.782	0.739	0.711	0.784	0.740	0.708	0.782
GNB	0.686	0.393	0.403	0.264	0.312	0.324	0.163	0.255	0.283

Number of authors	Precision			Recall			F Score		
Feature sets	F1	F2	F3	F1	F2	F3	F1	F2	F3
KNN-5	0.344	0.196	0.278	0.326	0.174	0.258	0.321	0.167	0.246
KNN-10	0.371	0.234	0.311	0.341	0.199	0.311	0.332	0.187	0.262
KNN-15	0.345	0.245	0.332	0.326	0.202	0.332	0.314	0.190	0.274
SVC	0.430	0.399	0.480	0.421	0.384	0.480	0.416	0.382	0.462
Linear SVC	0.488	0.302	0.429	0.484	0.300	0.429	0.481	0.298	0.426
GNB	0.017	0.058	0.047	0.041	0.058	0.047	0.008	0.022	0.040

Number of authors	Precision			Recall			F Score		
Feature sets	F1	F2	F3	F1	F2	F3	F1	F2	F3
KNN-5	0.313	0.164	0.259	0.281	0.122	0.198	0.281	0.118	0.195
KNN-10	0.311	0.159	0.264	0.293	0.133	0.219	0.286	0.125	0.214
KNN-15	0.320	0.216	0.304	0.299	0.147	0.234	0.292	0.140	0.229
SVC	0.394	0.354	0.470	0.387	0.333	0.442	0.382	0.331	0.444
Linear SVC	0.401	0.272	0.390	0.412	0.262	0.380	0.402	0.262	0.381
GNB	0.010	0.022	0.039	0.026	0.030	0.037	0.008	0.011	0.014

Out of all the classifiers the GNB performed the worst getting constant low scores however this should have been expected since most lexical distributions are not Gaussian.

A final observation of the data is that of the performance of the KNN classifier performing well only with a low amount of authors however the largest value of K (15) tended to provide the best results more often than not. This raises the question, how would the results change with increasingly higher values of K.

5. FUTURE WORK

To improve upon this work the first possibility would be to expand the feature set being used, this can be accomplished by expanding on the word based features such as more measures of vocabulary richness as mentioned in [6].

Experiments could also be run comparing the current and future results, with results obtained by from feature sets made of character n-grams and results obtained from combinations of n-gram feature sets alongside other feature sets. Work such as [9] could be used as inspiration especially since it used the same data set.

Yet another way to improve feature sets would be to include more semantic information such as frequencies of POS and

other information obtained from NLP processing of the text documents.

Finally the effects of normalization on the feature vectors can be tested. This step for many authors may involve a lot of processing and the time taken to preform it may out way its potential benefits.

Research into ensemble system may also be beneficial to reduce the task of the classifier and the potential of it making mistakes several classifiers could be train and act as sifts the first narrowing the author down to a small group of very similar authors, in terms of writing style, the second would then be able to identify the exact author. Alongside such a multi stage approach it may also be beneficial to include several differently trained models which could then all make a prediction and perform majority or weighted voting to determine the correct author.

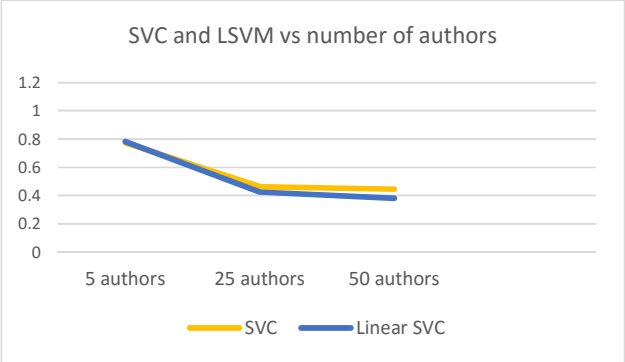

Figure 2 SVC and Linear SVC

6. CONCLUSION

From this work we have shown that given a small set of authors it is possible to perform author identification to a reasonable degree of accuracy using only lexical features with the best result being that of the linear SVM with 5 authors giving an f score of 0.782, however as the number of potential authors grows the scores drop to unacceptable levels. Therefore to make accurate classifiers which can handle large volumes of authors it will be necessary to use more that character and word based features.

Another surprising conclusion was that the combination of the two feature sets used (F3), in most cases, yielded a worse result than the use of only character based feature set (F1). This was unexpected and may indicate at either something wrong in the way the features were calculated or that some portion of the feature set is not helpful and may be acting as a red herring, making the classifier output incorrect classifications.

7. REFERENCES

[1] O. de Vel, A. Anderson, M. Corney and G. Mohay, "Mining e-Mail Content for Author Identification Forensics," *SIGMOD Rec.,* vol. 30, no. December 2001, pp. 55--64, 2001.

[2] A. a. P. H. a. G. N. Z. a. B. J. a. S. E. a. S. E. C. R. a. S. D. Narayanan, "On the Feasibility of Internet-Scale Author Identification," in *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, Washington, DC, USA, IEEE Computer Society, 2012, pp. 300--314.

[3] E. Stamatatos, "A Survey of Modern Authorship Attribution Methods," *J. Am. Soc. Inf. Sci. Technol.,* vol. 60, no. March 2009, pp. 538--556, 2009.

[4] D. M. a. A. G. a. D. D. L. a. E. G. D. D. L. a. S. A. a. D. F. a. L. Y. a. D. D. L. Consulting, "Author Identification on the Large Scale," in *In Proc. of the Meeting of the Classification Society of North America*, 2005.

[5] A. Abbasi and H. Chen, "Writeprints: A Stylometric Approach to Identity-level Identification and Similarity Detection in Cyberspace," *ACM Trans. Inf. Syst.,* vol. 26, no. March 2008, pp. 7:1--7:29, 2008.

[6] R. a. L. J. a. C. H. a. H. Z. Zheng, "A Framework for Authorship Identification of Online Messages: Writing-style Features and Classification Techniques," *J. Am. Soc. Inf. Sci. Technol.,* vol. 57, no. February 2006, pp. 378--393, 2006.

[7] ZhiLiu, "Reuter_50_50 Data Set," National Engineering Research Center For E-Learning Technology, 08 09 2011. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Reuter_50_50. [Accessed 05 January 2016].

[8] scikit-learn.org, "scikit-learn," [Online]. Available: http://scikit-learn.org/stable/index.html. [Accessed 05 01 2016].

[9] J. a. S. E. Houvardas, "N-Gram Feature Selection for Authorship Identification," in *Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, Berlin, Heidelberg, Springer-Verlag, 2006, pp. 77--86.