AUTHOR IDENTIFICATION BASED ON A HYBRID FEATURES SET USING MACHINE LEARNING TECHNIQUES

Samira Bourib
Speech Communication and Signal Processing Laboratory
Faculty of Electrical Engineering, BP 32, Bab Ezzouar
bouribsamira@gmail.com

DOI: 10.5281/zenodo.7492985

***Abstract*— This study explores a series of author attribution approaches that use some techniques for improving performances. Author attribution is the identification of the author of a text. It has attracted many attentions because of its relevance in a wide range of applications, including forensic investigations and plagiarism detection. A large number of features and approaches have been applied to this task. However, there has been a lack of studies that involve multiple datasets or that use a different range of classifications. Therefore, in this work, we explore several machine learning tools, the fusion between classifiers and different types of features that have been applied on the C50 database.**

Keywords— *Machine learning, MLP, SVM, LR, NLP Features, C50 Corpus, Word-Ngram, Stylometry.*

1 Introduction

Plagiarism detection can become ineffective when one does not have access to source documents, or when faced with a space as vast as the web with the increasing number of electronic documents as is often the case in electronic documents, or as is often the case with anti-plagiarism software. As a result, the development of tools for automatic text analysis and processing, and automatic text classification, has become important. Author identification techniques are one of the best known methods of text classification and have been widely used around the world to identify actual authors of anonymous texts. In recent times, there has been an increase in literary theft, loss of literary rights and concealment of the original author of a particular article or document. Anyone can take a copy of someone else's work and put it on a website or in an article with their name on it.

Also, verifying authorship is to assess whether the text in a dispute was written by the same author of the known texts or not (Sayoud 2012). This would be perfect if there was a system capable of analyzing and discovering the unstructured article to attribute the text to its original author.

Therefore, machine learning emerged to analyze the documents and extract features to predict the true author attribution. In this project, we conduct a research work to develop automatic text classification methods for author attribution (Juola 2006) (Stamatatos 1999) based on the linguistic style of a text written in English. We apply a machine learning approach by implementing a set of models that have proven their efficiency in different works dealing with document classification for author attribution, and then compare these approaches in order to find the best method for this task.

2 Description of the Corpus

For the second part of our experimental study we used the corpus named "C50" (Zhi Liu 2011), this database presents 50 authors each having 100 different texts, 50 texts for the test and 50 texts for the train, these texts have a size ranging from 28 to 95 sentences, 100 to 1000 words, 1691 to 6392 characters and from 1 to 3 A4 pages.

The documents are written in English and appear well written. For this corpus our experiments were done by increasing the data set and increasing the number of authors. The following table illustrates the method used:

Table 1: Size of the segments of the 'C50' corpus used in the experiments.

Dataset	Number of texts for TEST	Number of texts for TRAIN
Author 1 to author 10 (A1_A10)	50	500
Author 1 to author 20 (A1_A20)	100	1000
Author 1 to author 30 (A1_A30)	150	1500
Author 1 to author 40 (A1_A40)	200	2000
Author 1 to author 50 (A1_A50)	250	2500

3. Experiments of Author attribution

The objectives of the experiments carried out on our database are intended to evaluate the performances of the classification algorithms we have used. Our author attribution system is based on the following descriptors:

- Rare Word: words that appear only once or twice per document.
- Character N Grams: Groups of 2, 3, or 4 successive characters, with a sliding window.
- Words N-Grams: Groups of 1, 2, 3 or 4 successive words, with a sliding window.

Several approaches are considered such as SVM, MLP, Naïve bayes, KNN and Linear Regression. We segmented the corpus into two segments, the first one is used for training which contains 70% of the data, and the second segment is used for testing with 30% of the data.

3.1 Experimental algorithm and protocol

Setup: Begin
Choose a Classifier: 1-SVM, 2-MLP, 3-LR, 4-Stam, 5-Manh
Setup: End
Features: Begin
Read Tex File
Extract the Words
Extract the Characters
Concatenate Characters into Character-3Gram
Features:
End
Training: Begin
For every known Author (j=1…N):
Do Features Extraction
Build the Author Model
End For
Training: End
Testing: Begin
For every unknown author:
Do Features Extraction
Make an A.A. Classification
Determine the most probable Author (Aj)
End For
Testing: End
Result: Begin
Display Result (Aj)
Display the used Classifier
Display the used Feature
Result: End

Having carried out the steps listed above, we now proceed with the last step which is the evaluation of our models and the interpretation of the results. To make this possible we calculate the correct attribution rate of each model with the Authorship Attribution Score:

$$AAScore = \frac{\text{Number of texts well attributed}}{\text{Number of tested documents}} \quad (1)$$

$$\text{Error rate} = \frac{\text{Total texts} - \text{number of well assigned texts}}{\text{Total texts}} \quad (2)$$

3.2 Author attribution: Experiment part I

The objectives of the experiments carried out on our database are to evaluate the performances of the classification algorithms we have used. The work carried out is divided into three parts: the main algorithm, the evaluation of the results according to mathematical/statistical formulas and finally the interpretation of the results for the C50 corpus.

3.2.1 Results obtained with the MLP

The following figure shows the performance of the different features with the MLP classification method during the experiments. On the dataset there will be a part demonstrating the AAScore and another one for the error rate of author identification. In this histogram, we notice that even if we use the same method (MLP for this experiment) the features behave differently such that some are more accurate than others. In this experiment the feature Word is the most adapted for small datasets so that the AAScore reaches 90% of correct attribution for A1-A10 (a total of 50 texts 5 texts for each of the 10 authors), but by increasing the number of texts we notice that the percentage of error rate increases up to 89% and that the AAScore decreases down to 11% for the A1-A50.

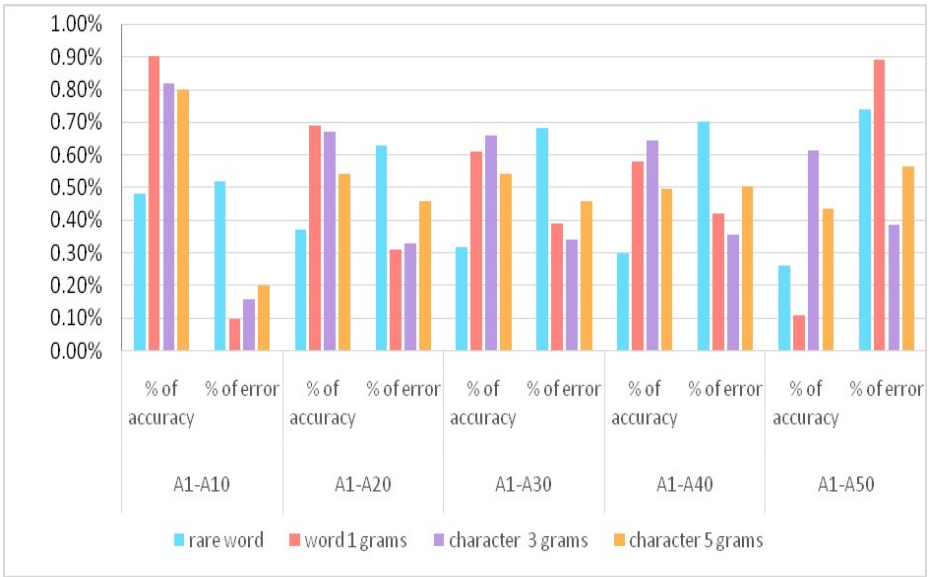

Figure 1. Author attribution score and error rate for each dataset using MLP with different features

The character-3grams and character-5grams appear to provide different performances. One can see very quickly that character-3grams are much more efficient than character-5grams. The character-3gram feature, compared to all other characteristics performs better when we increase the dataset so that the AAScore has reached 61.20% for the A1-A50. Finally, for the rare word feature, it is the least accurate with both small and large datasets: for example, its AAScore with a dataset of 50 texts (A1-A10) reached 48%. We conclude from this experiment using the MLP method that the word feature is the most suitable for small data sets and that character-3gram is the most or at least the best in large datasets.

3.2.2. Results obtained with the SVM

The following figure shows the performance of the different features when increasing the dataset size with the SVM classification method (SMO) in the task of author identification. In this figure, for each feature as a function of the dataset size there will be a part showing the AAScore and another one for the error rate. We notice that even if we use the same method for all features they behave differently and some are more accurate than others. In this experiment the feature word and character-3grams are very close. These two descriptors are the most suitable features for both small and large data sets so that the AAScore reaches 82% and 80% with error percentages of 18% and 20% for A1-A10 (a total of 50 texts 5 texts for each of the 10 authors) and 58% and 55.60% error percentages for the dataset A1-A50. Concerning the character-5gram it is adapted for small datasets it has almost the same performances as the features word and character-3grams with an AAscore that reaches 80% in A1-A10. But by increasing the dataset we notice that it is not adapted to manage a large number of texts because the error rate increases and reaches 53% and the AAScore reaches 47%. Finally, for the rare word, it is the least precise for the A1-A10 dataset: it reaches 60% but the more we increase the dataset the less precise becomes the AA classification and the more the error rate increases, reaching 74% for the A1-A50.

We conclude, from this experiment using the SVM classifier, that the Word feature and character-3grams are the best adapted for small and large data sets even if their scores are quite low. And for character5Grams and rare words they are good for small datasets but have a high error rate for large datasets.

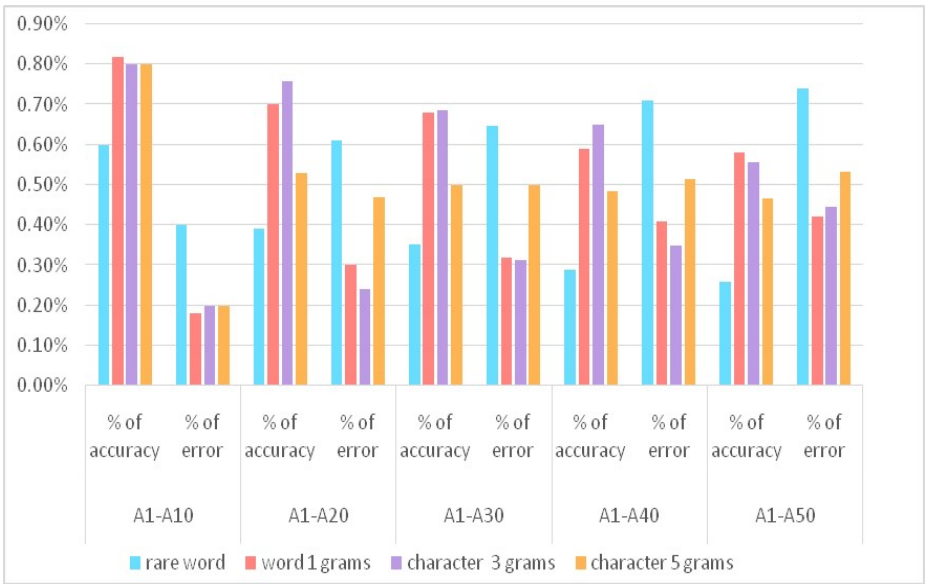

Figure 2. AAScore and error rate for each dataset using the SVM with the different features

3.2.3. Results obtained with Linear Regression

The following figure shows the performances of the different features as a function of the dataset size with the LR classification method during the author identification experiments.
In this experiment the features Word and character-3grams have very close performances. These two features are the most suitable features for both small and large data sets so that the AAScore reaches 80% and 74%, with error rates of 20% and 26% with A1-A10, and an AAScore of 52.80% and 62.00% and error rates of 47.20% and 38% with A1-A50. Comparing the two types, the character-3Grams is more accurate. The character-5grams is suitable for small datasets and is almost similar to the features Word and character-3grams with an AAscore that reaches 76% with A1-A10. But by increasing the dataset, we notice that it is not adapted to manage a large number of texts because the error rates increase and reach 51.60%. Finally, for the rare word feature it is the least precise for the A1-A10 dataset: it reaches 44% which is already low and the more we increase the dataset the less precise becomes the classification and the more the error rate increases and reaches 76% with A1-A50.
We conclude, from this experiment using the LR method, that the features Word and character-3grams are the most suitable for small and large datasets even if their scores are quite low. The character-5Grams is good for small datasets but has a high error rate for large datasets. Finally, for the rare word, the results are bad in small and large data sets.

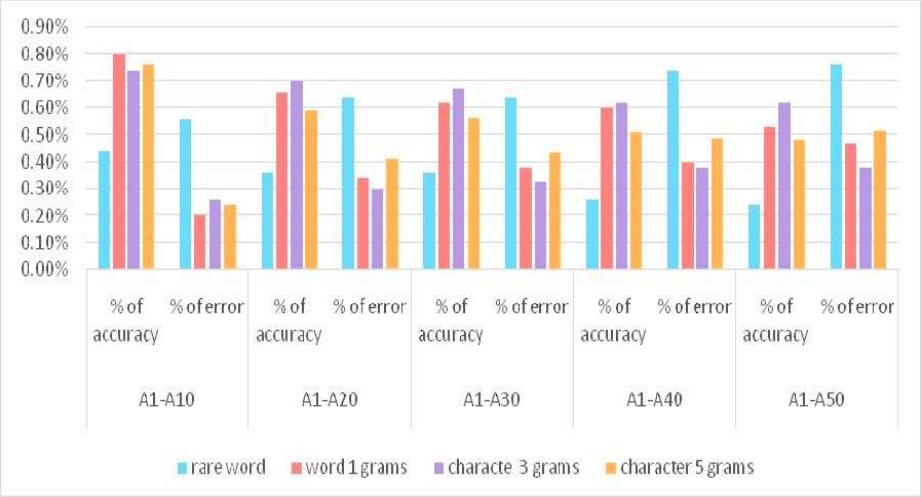

Figure 3. AAScore and error rate for each dataset using the LR with the different features

3.2.4 Average robustness

In order to evaluate the performances of the attribution methods and the features, we defined a specific ratio, which we called "Average robustness or accuracy". This ratio is calculated as follows:

$$\text{Average robustness} = \frac{\sum_{j=1}^{n} \text{AAScore of the characteristics of the method used}}{4} \quad (3)$$

Through the following histogram we will discover which method of author attribution is the most robust. Example: let us consider the following table for the different rsults of the MLP.
The average robustness or accuracy is equal to **(90+48+84+80)/4=75.5%.**

Table 2: Average robustness for the different features using MLP

MLP (A1-A10)			
Word 1Grams	Rare Word	Character 3Grams	Character 5Grams
90.00%	48.00%	84.00%	80.00%

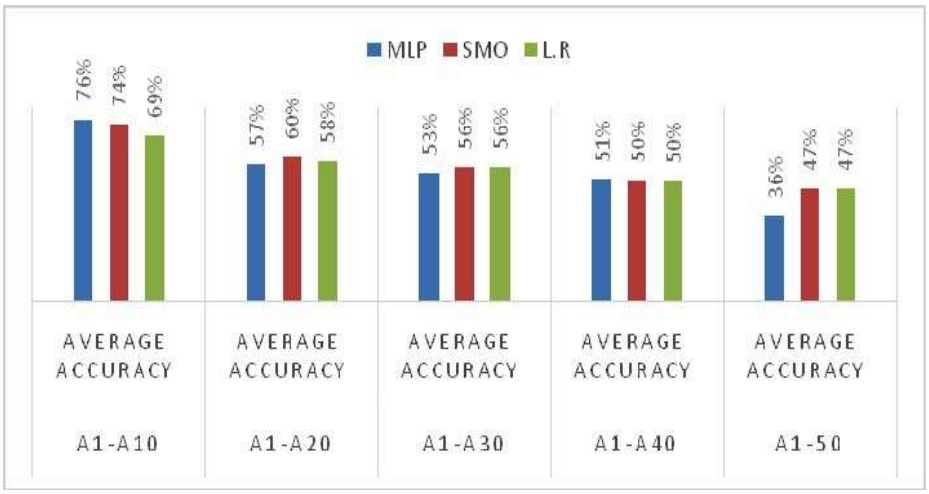

Figure 4. Average robustness for each size of dataset

We notice that for the same size of dataset, the three methods offer more or less similar results. Let us take the example of A1-A10, the MLP, SVM and LR scores are equal to 75.50%, 73.50% and 68.50% respectively, and this remark also applies to larger datasets. But it appears that some methods are more robust than others, for example the MLP is more accurate for small datasets but not for large datasets.

3.3 Author attribution: Experiment Part II

In table (3), we can make several interesting observations: regarding the features, we can notice that the best accuracy was obtained by character tetragrams (average score of 89.02%), which consist of a combination of four characters. The worst accuracy was obtained by using rare words (average score of 59.54%). Also, regarding the classifiers, it can be noticed that the best accuracy was obtained by SMO-SVM (average score of 89.16%), while the worst result was obtained for centroid Stamatatos distance (average score of 58.19%), which is probably due to the small training dataset. The best score obtained in these experiments, is about 99%, obtained with SMO-SVM and word-bigrams, showing that the author attribution on our subset (split C50 dataset) could be done quite well without exceeding 2% of attribution error.
Table (3) can be summarised as an overall attribution score according to the 2 main types of features, namely: character-based features and word-based features. Here, it can be clearly observed that word-based features are better than character-based features with only 4 best scores (in bold blue) and 4 best scores (in bold red), as shown in table (3).

Table 3: Good author attribution scores.

Feature / Classifier	Character Bigram	Character Tri-gram	Character Tetra gram	Word	Word Bi gram	Word Tri gram	Word Tetra gram	Rare Word	Average Accuracy per classifier(%)
MLP	87.77	91.11	95.55	73.33	73.70	70.22	69.9	39.7	75.16
Naïve Bayes	88.88	84.44	90	56.66	71.11	30	32.33	56.66	58.62
SMO-SVM	83.33	90	91.11	97.77	98.88	92.22	83.33	76.66	89.16
KNN	61.11	87.77	94.44	95.55	96.66	95.55	87.77	61.11	85.00
Linear Regression	83.33	88.88	93.33	97.77	92.22	83.33	66.66	61.11	83.33
Manhattan Distance	86.66	83.33	86.66	96.66	95.55	62.22	36.66	65.55	65.83
Stamatatos Distance	52.22	60	67.77	82.22	64.44	51.11	45.55	42.22	58.19
Cosine Distance	84.44	91.11	93.33	90	93.33	54.44	45.55	73.33	78.19
Average Accuracy per feature(%)	78.46	84.58	89.02	86.24	85.73	67.38	58.46	59.54	

Table 4: The fusion score of different characters Ngram and Words N-gram for author attribution with the SMO-SVM classifier.

Feature used / Classifier used	Fusion Character 2 3 4	Character 2 word 2 3	Character 2 3 Word 2	Char 3 4 word 2	Word 2 char 2 3	Word 3 char 2 3	Word 4 char 2 3	Word 2 4 char 2	Word 2 4 char 3	Word 2 4 char 4	Word 1 2 3	Word 1 2 4
Score in %	88.88	98.88	91.11	93.33	97.77	98.88	98.88	95.55	97.77	97.77	98.88	98.88

Table 5: Score of good Author Attribution of the SMO-SVM classifier.

Feature used	Character Bigram	Character Tri-gram	Character Tetra-gram	Word	Word Bi-gram	Word Tri-gram	Word Tetra-gram	Rare Word	Average Accuracy per classifier
Score in %	83.33	90	91.11	97.77	98.88	92.22	83.33	76.66	89.16

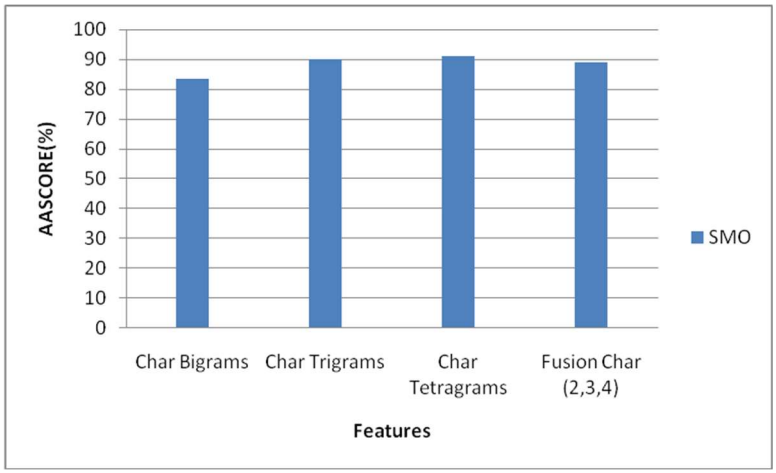

Figure 5. Score of fusion with the different characters Ngrams using SMO-SVM

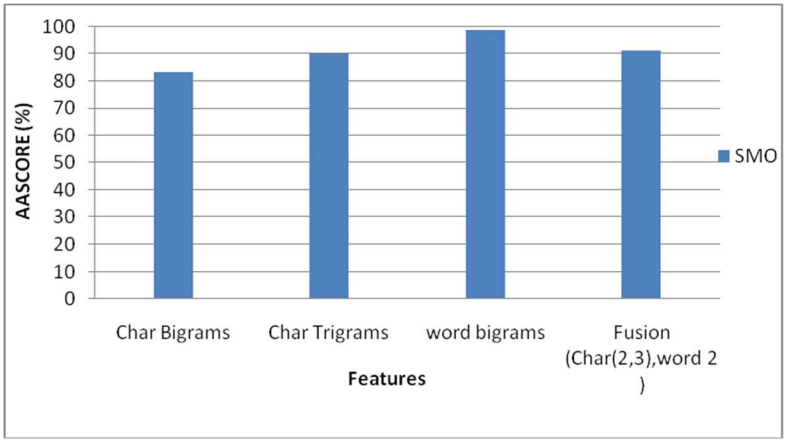

Figure 6. Score of fusion with the different characters Ngrams using SMO-SVM and word 2gram using SMO-SVM

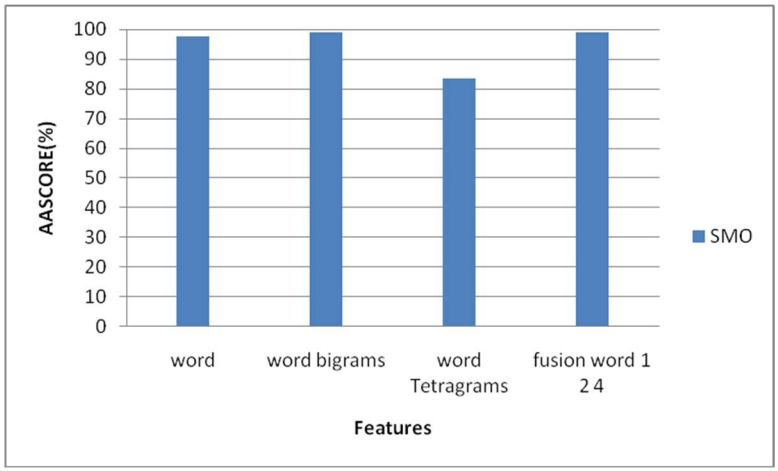

Figure 7. Score of fusion with the different word Ngrams using SMO-SVM

4. Conclusion

The work that we have presented is concerned with the automatic classification of textual documents according to the author. It is getting more and more interest in the scientific community as it provides interesting solutions to many real problems. The corpus we used to validate our experiments, and which belongs to Reuters_50_50, is built around a database consisting of 50 authors for training and 50 authors for testing, each of which has 50 different texts with an average size of about 100 to 1000 words per text.

Generally, the process of automatic classification consists of two steps: first, the pre-processing step and the representation of text in vector form adapted to the learning algorithms knowing that there is no learning method capable of directly processing unstructured data such as texts. Secondly, the choice of a learning method that gives better results with a good accuracy rate, and the method is fast for processing large data.

During our experiments, we investigated different features, based on characters and words, and classifiers such as SVM-SMO, MLP, K-NND and Linear Regression Naive bays, etc.

This study led us first to conclude that the categorization of textual patterns is very different from the categorization of other patterns which makes the task of morphological analysis very complex. Secondly, after analyzing the results obtained, it was found that the classification scores were different depending on the association of the feature with the classifier.

We also presented the different approaches and the various types of combination of classifiers as well as the different levels of fusion that exist in the literature related to the fusion domain, where we saw the importance of such a fusion.

In order to improve the performance, as perspectives, we propose the following tasks: applying other approaches of text representation (ending gram for example), implementing other classifiers (such as deep learning: CNN, RNN ...), using other distances (cosine, correlation, etc.), and using ontologies or dictionaries to enrich the data representation.

References

P. Juola., 2006 JGAAP : Authorship Identification. Foundations and Trends in Information Retrieval Vol. 1, No. 3 (2006) 233–334, Now Publisher.

H. Sayoud, 2012. "Author Discrimination between the Holy Quran and Prophet's Statements", LLC journal, Literary and Linguistic Computing Journal, Oxford-University Press. Accepted and published on-line in 2012. Citation reference: doi: 10.1093/llc/fqs014, Volume 27,pp427-444, Nov,2012.

E. Stamatatos, N. Fakotakis, and G. Kokkinakis, 1999. Automatic Authorship identification In Proc. of the 9th Conf. of the European Chapter of the Association for Computational Linguistics, pp. 158-164, 1999.

Z. Liu, 2011. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/ datasets/Reuter_50_50. Donated in 2011.