# Authorship Identification from Document Features

**Xiyu Chen (xc163)**
**Zhengrong Gu (zg120)**

## Abstract

Authorship identification has been a hot topic especially in the Internet age. Following previous work in literature, we found strong interest in finding the most helpful document characteristics for identifying authorship. In this work, we implemented logistic regression model and Support Vector Machine (SVM) on features such as structure features, frequent words, function words, n-grams.

## 1 Introduction

Authorship identification is a problem with long history and wide range of application. It is a process of identifying authors of documents based on examples of authors' works. In our data-rich age, with more data and more advanced analysis techniques available, the topic of authorship identification has attracted more attention in academics. Datasets such as Reuters Corpus Volume I (RCV1) available by Reuters, Ltd, Gutenberg Dataset from Project Gutenberg website, have been widely used in authorship identification related research. Since these datasets are very unique, voluminous and informative, analyzing them would be a very interesting task. Through researching in analyzing these datasets, academics get motivated by exiting methods that are widely used in other areas and invent unique methods for specific problem in authorship identification. We are interested in those methods that well performed for authorship identification topic. Specifically, interested in finding the features of documents that are most effective in helping identify authors. In this paper, by using Support Vector Machine (SVM) and logistic regression, we explore the n-gram features and other structural level features in the topic of authorship identification.

## 2 Background

Reuters Corpus Volume I (RCV1) is an archive of over 800,000 manually categorized stories made available by Reuters, Ltd. for research purpose. The necessary documentation of this dataset can be found in Lewis et al. (2004). This documentation describes RCV1's operational setting, especially the categories and how they are assigned. It also addresses the semantics of the category assignments and provides the insight into operational text categorization. Besides, it introduces RCV1-v2, a new version of dataset comparing to RCV1-v1.

Academics pay special attention to methods that can solve authorship identification problem. Speaking of models, in Madigan et al. (2005), authors focus on multinomial or polytomous generalization of logistic regression. They used Bayesian multinomial logistic regression with Laplace prior to building classifiers. The output is an important advantage of this method as it is an estimate of probability rather than documents belonging to each of the possible classes. Speaking of features, most of the previous work mainly explores stylistic and linguistic features. Iyer et al. (2019) explored with combination of features, including clubbing stylometric meta features like bi-grams, POS bigrams and word/POS pairs. And the accuracy increases significantly after introducing bi-grams. Zhang et al. (2014) explores the abstract semantic patterns of sentences and creates 10 features with both semantic and non-semantic features. With regard to texts, there is related work focusing on unstructured or structured texts. For example, Zhang et al. (2014) researched on the unstructured dataset.

# 3 Data

## 3.1 Data Source

We searched datasets that are mostly used for authorship identification and found the RCV1 is a perfect fit to our research project. We selected the subset of RCV1, named Reuter_50_50 dataset. This corpus has already been used in some authorship identification experiments. It includes top 50 authors, whose total size of articles are the largest. According to the UCI Machine Learning Repository, considering minimizing the topic factor in distinguishing among the texts, 50 authors of texts labeled with at least one subtopic of the class CCAT (corporate/industrial) are selected. The training corpus consists 2500 texts and the text corpus consists another 2500 texts. Because each author has the same number of texts included in the corpus, this dataset is very balanced and is very useful in pattern recognition, classification and clustering tasks.

## 3.2 Data Preprocessing

Data preprocessing is an important and necessary preparation step. Since mistakes, redundancies, missing values and inconsistencies would all compromise the integrity of the data set, we need to preprocess the data set before training an algorithm.

In Natural Language Processing, one of the most common steps in text preprocessing is removing stop words. These words are actually the most common words in any language and does not add much meaning to the text. In this research, since the texts are all written in English and there are a lot of stop words in those texts, we remove stop words at the first step of data preprocessing.

Stemming is also necessary. Stemming is a process for removing the commoner morphological endings from words in English. We used PorterStemmer() from NLTK package to realize this.

Another step is reducing all letters to lower case. It allows instances of Apple at the beginning of a sentence to match with a query of apple.

# 4 Methodology

## 4.1 Feature Construction

### 4.1.1 Structure Features

The token structure is focused in our feature construction step. The tokens' mean, maximum, minimum, 10th percentile, 20th percentile, 80th percentile, and 90th percentile lengths are chosen as the structure features, see Figure 1-4.

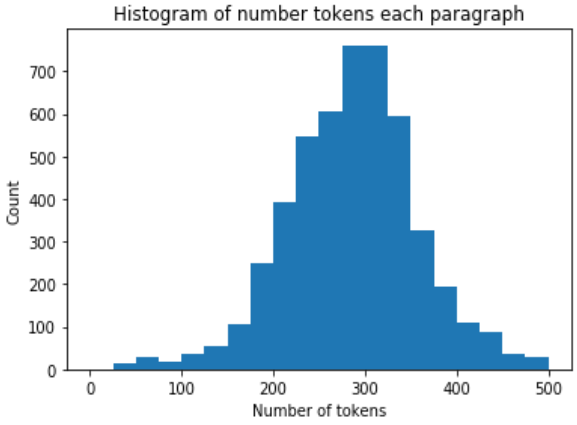


Figure 1: Histogram of number of tokens each paragraph.

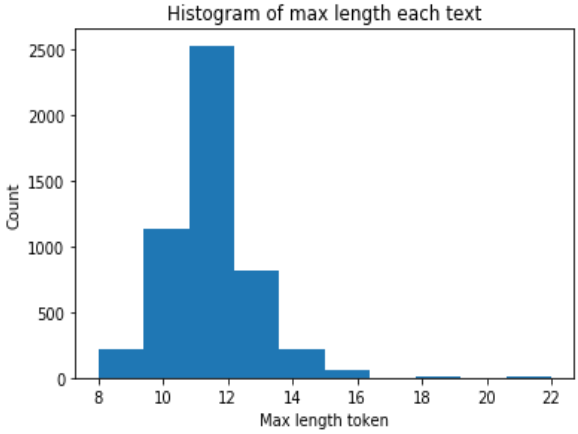


Figure 2: Histogram of maximum length of tokens each paragraph.

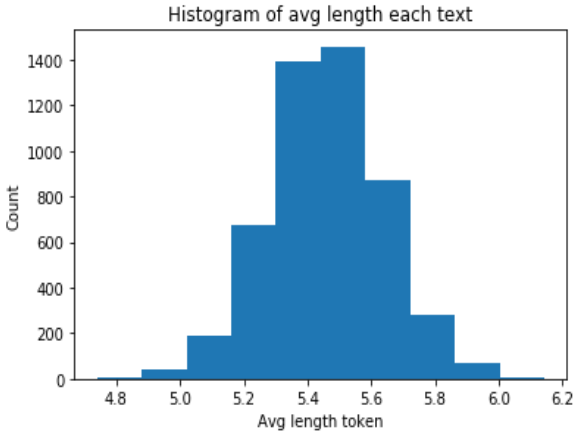


Figure 3: Histogram of average length of tokens each paragraph.

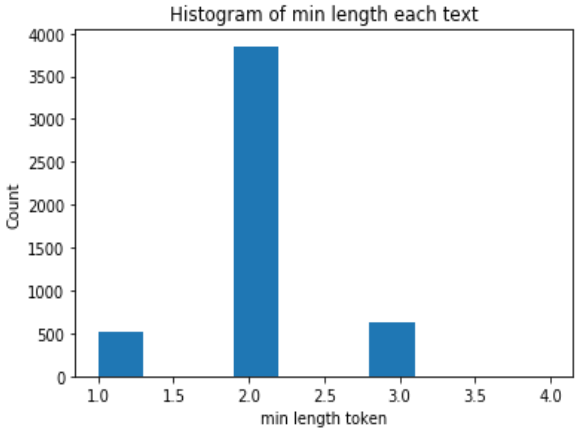


Figure 4: Histogram of minimum length of tokens each paragraph.

#### 4.1.2 Lexical Level Features

Function words are chosen to be the lexical level features. Function words are chosen because it is content independent. Because of this, function words can imply authors' preference of word usages and writing styles. There are 176 function words and the count of them are used as the features in the model.

#### 4.1.3 N-gram Vectorization

N-gram is explored in the study. Analyzing n-gram is a commonly used methodology in authorship identification. One disadvantage of this method, however is that, it can be influenced by the topics and genres of the authors' writings. Thus, we want to explore the influence of the different N choices.

### 4.2 Algorithm

After the feature sets are constructed, two predictive methods are used to train the final model: logistic regression and Support Vector Machine (SVM).

#### 4.2.1 Logistic Regression

A logistic regression model is implemented on the function word features and the structure feature sets as mentioned in the section 4.1. Those two feature sets have lower dimensionality than the n-gram features. Logistic regression is an ideal model for low-dimensional dataset. The results of this model are in the section 5.

#### 4.2.2 Support Vector Machine

Support vector machine is applied to function word features, structure feature sets and n-gram features. We implemented SVM model to all these three features because SVM can deal with high-dimensional input well and does not require feature selection. The results can be found in section 5.

## 5 Results

| | **Train** | **Test** |
|---|---|---|
| Structure | 0.0905 | 0.0706 |
| Function words | 0.341 | 0.200 |

Table 1: Results of logistic regression.

The logistic regression results are shown in the Table 1. The structure features achieve a training accuracy of 0.0905 and test accuracy of 0.0706, which is slightly better than the base line. The function word achieves a training accuracy of 0.341 and test accuracy of 0.200. It is obviously better than the structure features', but there is big difference between training accuracy and test accuracy.

| | **Train** | **Test** |
|---|---|---|
| Structure | 0.0851 | 0.0526 |
| Function words | 0.426 | 0.205 |
| Frequent tokens | 0.936 | 0.747 |
| 2-gram | 0.847 | 0.625 |
| 3-gram | 0.743 | 0.478 |

Table 2: Results of SVM.

The SVM results are shown in the above Table 2. Frequent tokens work best with 0.936 training accuracy and 0.747 test accuracy. Then is the 2-gram, 0.847 training accuracy and 0.625 test accuracy. 3-gram has 0.743 training accuracy and 0.478 test accuracy. Those three are way better than

the function words with training accuracy 0.426 and test accuracy 0.205. The worst model is the one using structure features. Its training accuracy is 0.0851 and test accuracy is 0.0526.

Overall, both in Logistic regression and SVM models, the test accuracy is all above the baseline which is 0.02 by just random guessing. We can say that features we selected—structure features, functions words, frequent tokens, 2-gram, 3-gram are very useful in identifying authorship. However, high accuracy may stem from other features that are 'hidden' in these features. Because of the limitation of the research, we cannot fully filter out features such as topics of works from n-grams and frequent words.

## 6 Conclusion

Using logistic regression model and SVM model can attain test accuracy that is higher than the 0.02 baseline. We can say that these two models are useful in authorship identification.

Different features perform very differently. In the research, structure features perform worst while frequent words perform best. n-gram also have a good performance.

Though we have prepared a very balanced dataset which is very suitable for authorship identification task, we have preprocessed the data set carefully and we designed the methodology part based on insights from previous literature, there are still limitations in our research. One limitation is that, we are not able to filter out the effects of text topics. The topics of authors' writings would decrease the generalization of the frequent words and n-gram models. This idea is consistent with our findings that 3-gram model performs worse than 2-gram model, because when n increases, the topic 'dilutes'.

## References


David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research 5: 361-397. https://www.jmlr.org/papers/volume5/lewis04a/lewis04a.pdf*

David Madigan, Alexander Genkin, David D. Lewis, and Dmitry Fradkin. 2005. Bayesian Multinomial Logistic Regression for Author Identification. *AIP Conference Proceedings, 803, pp.509-516. https://aip.scitation.org/doi/abs/10.1063/1.2149832*

Rahul Radhakrishnan Iyer, Carolyn Penstein Rose. A Machine Learning Framework for Authorship Identification from Texts. 2019. *arXiv preprint: 1912.10204. https://arxiv.org/abs/1912.10204*

Chunxia Zhang, Xindong Wu, Zhendong Niu, and Wei Ding. 2014. Authorship Identification from Unstructured Texts. *Knowledge-Based Systems, volume 66, pages 99-111. https://www.sciencedirect.com/science/article/abs/pii/S0950705114001476?via%3Dihub*