De-anonymizing Review-Style Texts

Anonymous Author(s)
Affiliation
Address
email

Abstract

Authorship attribution, the process of inferring the author of an anonymous document based on writing style characteristics of the author, has long history and a wide range of applications. In this study, this technique, combined with two machine learning classifiers, has been used to de-anonymize the identity of reviewers of a seminar-style graduate class. some stylistic features from the set of reviews were extracted to help identify individual reviewers based solely on their writing style. The selected features were chosen from lexical, structural and idiosyncratic varieties to increase the classification accuracy. For learning the input features and predicting the reviewer identity of unlabeled reviews, random forest and logistic regression classifiers were used.

1 Introduction

Authorship attribution is referred to the task of mapping a group of anonymous documents to their corresponding authors after learning some distinguishing features from the writing style of that group of authors. The chosen features should be able to discriminate between the writing style of different authors. There exist a vast variety of stylistic features in the literature ranging from lexical, syntactic, structural, and idiosyncratic markers. Lexical features represent lexical variation of the document. These include word frequencies, average word or sentence-length, and vocabulary richness among many others. Syntactic features demonstrate syntactic patterns that an author tends to use more frequently, including sentence and phrase structure, part-of-speech, and function words. Structural features relate to the layout and organization of the document, i.e. the use of comments, bullets, braces,fonts, colors and so on. Idiosyncratic features include misspellings and grammatical mistakes.

Authorship attribution has a long history starting with the seminal work of [1] which attributed twelve disputed papers to the same author, and their results were verified by literature scholars. Their approach relied on Bayesian statistical analysis of frequencies of occurrence of a set of words. Since then there have been similar studies in the field, on different anonymous domains such as emails, blogs, and online forums [3], [4], [6], [5], [8]. Most of this research has been devoted to the evaluation of various effective stylistic features or the choice of model parameters that will contribute to the greater accuracy of the proposed methods.

In the simplest case, having a document of unknown authorship and a set of candidate authors for whom the labeled document samples are available, the task is to assign the unlabeled document to one candidate author. From the machine learning point of view, its the same as multi-class single-label text classification task [2]. In a more general form, there are multiple unlabeled documents for which their corresponding authors should be found. This task, which is the case for this work, falls within the multi-class multi label text classification category.

Although it can violate the privacy of individuals in some cases, authorship attribution can have a number of useful applications as well, including but not limited to the followings:

- Author Authenticity Verification: determining whether a document has been written by the claimed author.

- Plagiarism Detection: detecting if a document has been Plagiarized.

- Author Profile Extraction: extracting some specific characteristics,(i.e. age or gender), about the author of a document.

- Collaborative Writing Detection: determining whether the text has been written by more than one individual.

- Document Age Estimation: Estimating how old a piece of text is.

However, in some cases, anonymous authorship attribution is done mainly to indulge the curiosity of the analyst. One can think of many of such scenarios, among which is conference-review de-anonymization, which was introduced by [6] in 2011. Assuming a conference community member has collected sufficient unblinded(labeled) review samples from his colleagues over time by serving on different conferences, the task is to deduce the identity of the authors for unlabeled review samples. One characteristic of this specific scenario is the short length of reviews and their more limited and technical-oriented vocabulary.
In [6], the features are the frequency of author-specific unigrams, bigrams, and trigrams in a particular author's reviews. In stylometry, n-grams are defined to be the word sequences of length n. For their classification algorithm, they've used naive Bayes classifier.

2 Our Work

In this work, dataset consists of reviews of a seminar-style graduate class, submitted to a HotCRP system. It includes two hundred paper reviews for the total number of eleven reviewers.

2.1 Feature Selection

One difficulty in this work was extracting appropriate features to feed to classifier algorithms. One reason was limited length of reviews compared to average document length in routine text classification problems. Another reason was academic content of reviews, which caused all reviewers to focus on the same vocabulary simultaneously.
In addition, as overall number of review samples was only two hundred, number of features had to be selected very judiciously. The reason was to prevent over-fitting that can happen if number of features exceed number of data samples.
For selecting discriminating features, first of all some statistical data about the text (i.e. number of sentences, words, characters, etc.) was considered. Then, the text was tokenized into words in order to match it against some regular expressions which encoded the most frequent words and phrases. Some reviewers seemed to stick to a narrower set of vocabulary for approval or denial of an idea. These idiosyncratic vocabulary are usually an appropriate choice for discriminating features. So, the relevant vocabulary were placed in a bag of words and the frequency of such words were tallied separately. Some reviewers, on the other hand, were rather difficult to classify, as they didn't follow a similar writing style in their reviews. For identifying these authors, a combination of more sophisticated features had to be used. At this step and after removing stop words(i.e. the, and, etc.), frequency of bigrams were counted. Frequency of trigrams was not considered as a feature, as trigrams happened very rarely for a given review and did not help improving classification accuracy. Another set of useful discriminating features were obtained by analyzing the reviews structure. Most of reviewers followed a similar structural pattern in all of their reviews(i.e. comments in enumerated lists, or using special fixed words at the beginning of the review, etc.), so the frequency of such structural patterns were also taken into account. All mentioned features, represented as numerical values, formed the feature vector for a review sample. Feature selection code is written in Python and for some specific functionalities, such as counting the number of bigrams, wrapper functions of NLTK [6] were called.

2.2 Classifying the Reviewers

Having extracted some discriminating features, next step was to test the quality of our choice of features using classifier algorithms. Two distinct classifiers, a random forest classifier and a logistic regression classifier, were used for learning feature vectors of the labeled data. The reason for choosing two classifiers was to estimate the efficiency of feature extraction algorithm using different classification approaches. Specifically, random forest was of special interest to this task, as it randomly uses a fraction of features at each split. Logistic regression algorithm, has the characteristic of working well in settings where feature vectors are sparse, due to its special parameterization. As in this work, some features are more author-specific, by deploying logistic regression classifier, classification accuracy might be improved.

For a multi-label classifier, *accuracy* is defined to be the fraction of unlabeled samples that are labeled correctly by the classifier. In this work, classification's accuracy has been used as a measure to evaluate the efficiency of feature extraction algorithm.

2.3 results

In order to study the effect of dataset size on the accuracy of classification, the tests were performed on variable-size trainingtest sets. At each round, a different size of training set was used as the classifier's input, and the remaining samples were used as the test set. Training set was drawn uniformly at random from all two hundred reviews. However, it was checked to ensure that text samples for a particular reviewer were not be below a given threshold. This initial check wass set in order to avoid over-fitting on some particular person's reviews or under-fitting on the others. For the test set, the labels were anonymized using a preprocessing script that removes identity information from reviews. Then, the test set was given to classifiers in order to predict corresponding labels. Finally, accuracy of the predications was calculated by comparing them to true labels. Classification accuracy of this setting, by considering different ratio for training and test sets, are depicted in Table1(Results are rounded for clarity):

Table 1: Comparison of Classifiers' Accuracy

Training/Test Ratio	Random Forest	Logistic Regression
4	0.72	0.74
7/3	0.65	0.76
3/2	0.64	0.67
1	0.70	71

Figure 1, shows the accuracy of running classifiers on different ratio of training and test data. Not surprisingly, increasing the size of training set, yields to higher accuracy in classifiers' predictions. Increasing number of training samples can alleviate small size of reviews. (In this figure, *x-axis* denotes ratio of review samples allocated to training and test sets.)

Another experiment assumed a fixed ratio between training and test seta(80%) and studied the effect of test data increase on the classifiers' accuracy. So, at the beginning, the classifiers performed predictions only on a randomly-chosen subset of test set and accuracy results were recorded. In the next steps, test set size was increased per step and the experiment was repeated for that portion of test set. So, in these experiments, the remaining test data were not used.
As it can be seen in Figure 2., with more test samples, the rate of correct labeling improves. Thats because some reviewers' writing styles are more difficult to classify. Increasing test set size, mitigates such occasions that only those hard-to-classify reviews are chosen on the test set. The results for varying test set size can be seen in Figure 2. In this figure, *x-axis*, represents number of test set samples, assuming a fixed 157 review samples are used for training.

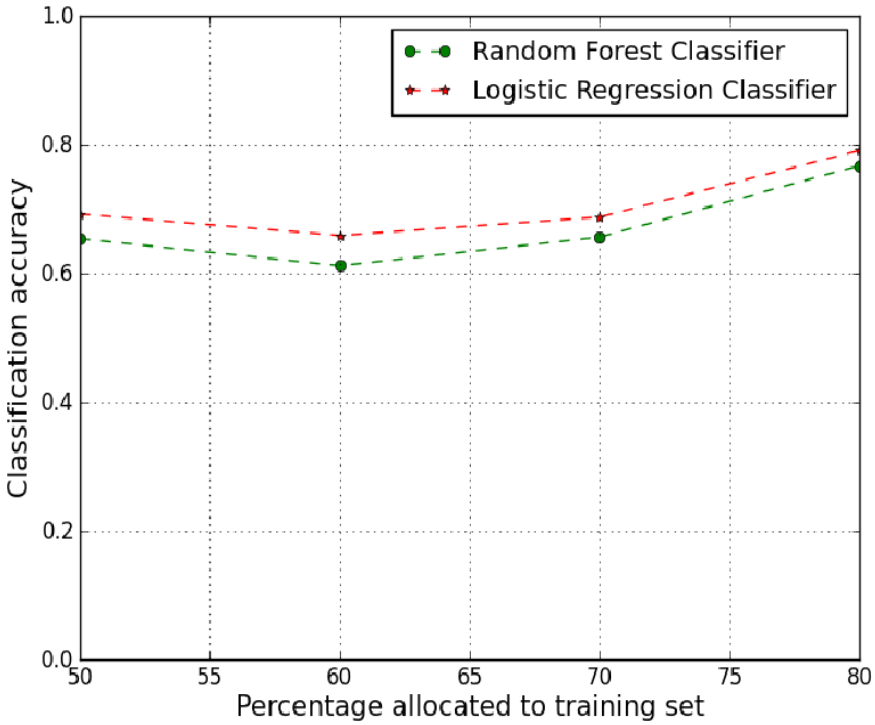

Figure 1: Classifier accuracy for different training set size

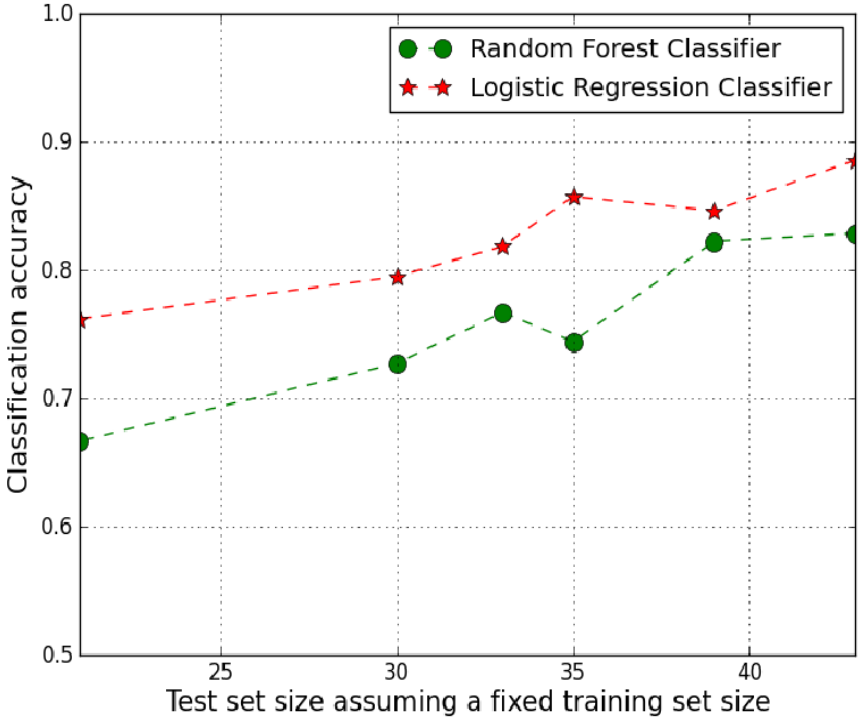

Figure 2: Classifier accuracy for different test set size and a fixed training size

3 Conclusion

In this paper, authorship attribution techniques are used to de-anonymize the identity of reviewers in a seminar-based graduate class. Various features from the texts are extracted and given to random forest and logistic regression classifiers and their accuracy of classification is compared. The results show that accuracy of the results is data-dependent and improves by increasing training set size. Although features were chosen very carefully, the results show that further analyzing the input texts is needed to improve feature extraction algorithm and increase classifiers accuracy.

References

[1] Mosteller, F. & Wallace, D.L. (1964). Inference and disputed authorship: The Federalist. Addison-Wesley.

[2] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1).

[3] Stamatatos, E., Fakotakis, N., Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. Computational Linguistics, 26(4), 471495, 2000.

[4] Keselj, V., Peng, F., Cercone, N., Thomas, C. (2003). N-gram-based author profiles for authorship attribution. In Proceedings of the Pacific Association for Computational Linguistics (pp. 255-264).

[5] Savoy, J. (2012). Authorship Attribution Based on Specific Vocabulary. ACM Transactions on Information Systems (TOIS), 30(2), 12.

[6] Nanavati, M., Taylor, N., Aiello, W., Warfield, A. (2011, August). Herbert WestDeanonymizer. In Proceedings of the 6th USENIX conference on Hot topics in security, ser. HotSec (Vol. 11, pp. 6-6).

[7] Zheng, R., Li, J., Chen, H., Huang, Z. (2006). A framework for authorship identification of online messages: Writing?style features and classification techniques. Journal of the American Society for Information Science and Technology, 57(3), 378-393.

[8] Pearl, L., Steyvers, M. (2012). Detecting authorship deception: a supervised machine learning approach using author writeprints. Literary and linguistic computing, 27(2), 183-196.

[9] Bird, S. (2006, July). NLTK: the natural language toolkit. In Proceedings of the COLING/ACL on Interactive presentation sessions (pp. 69-72). Association for Computational Linguistics.

[10] Almishari, M., Tsudik, G. (2012). Exploring Linkability of User Reviews. In Computer SecurityESORICS 2012 (pp. 307-324). Springer Berlin Heidelberg.