CMPE 544 - Fall 2022
Pattern Recognition Project Final Report
Authorship Attribution on Tweets

Elvan Karasu Ozan Özgür

1 Project Description

Authorship attribution is the process that aims to determine the author of a text. Common applications of authorship attribution involve finding the author of an important text when multiple writers are claiming to be the actual author, verifying the identity of the author and detection of malicious activities of writers who write under multiple names. Information for each text is represented as a set of features. After that, a classification or clustering model is used to predict an author. Reliability of authorship attribution methods largely depend on the amount of data. For example, projects working on long literary texts report higher metrics, compared to ones working on social media texts.

Authorship attribution can be studied by closed set and open set approaches. Closed set approach assumes that the authors coming in test time are known and learned in training time; whereas open set approach accepts unknown authors at test time. [1] The first part of this project was to solve this problem with closed set approach. Now we focus on the open set approach and present our results.

2 Method

2.1 Dataset

Data collection is tricky for authorship attribution task since it requires multiple tweets per user. For example, Rocha et. al do their experiments for this task with 50, 100, 200, and 500 tweets per each user, and about 50 users. [2] Considering these requirements, we chose Sentiment140 dataset. Sentiment140 is a dataset containing 1.6 million tweets. It is actually for sentiment classification; however it can be used for authorship attribution task as well. The top 50 users with the most number of tweets have larger than 150 tweets. This dataset has both plenty of data, and variability among users. [3]

The first task is to create a dataset from the texts of the possible authors, then use feature extraction methods. After that, a set of features are generated from each text.

Features related to word frequencies and n-grams are shown to be effective and reliable. [4] The dataset is formed such that it contains 50 users. There are 75, 35, and 20 tweets per user in the training, validation, and test sets, respectively. For the open set approach, 3157 tweets whose authors are different from the known authors are also added to the dataset. Open set approach contains unknown classes at the testing time. [1] Therefore; 3157 tweets, whose authors are different from the known authors, are added to the dataset.

2.2 Preprocessing

Times, dates, mentions, URLs are replaced with ... tokens in order to simplify and generalize these information. For example, if the tweet is “”, it is converted to ... . The users whose tweets are always the same are also removed from the dataset since they can be misleading in performance evaluation. For example, lost dog

2.3 Feature Extraction

Term Frequency-Inverse Document Frequency (TF-IDF) is used for feature extraction. [5] In the previous step of the project, character level 4-gram features was used since Rocha et. al found the best performance for n = 4 among different character level n-grams. [2] In general, they obtain even better performance when character level 4-gram features are used together with word level n-grams. [2] Therefore, word level 1-grams is also tried.

2.4 Model

In the first part of the project, the closed set approach was solved with multiclass classification. However, in open set, the problem is considered as a multiclass classification with a reject option. The model is expected to reject the samples whose authors are not in the training set. A new metric is used in order to observe the rejection performance, which is the ratio of samples that are rejected over the samples that should be rejected (coming from unknown authors).

In order to have a reject option, k-nearest neighbors (kNN) algorithm is modified such that it rejects the samples whose similarity are below a threshold. This threshold is chosen using the distributions coming from intraclass distances for each class. In the beginning, classification is the same as kNN. However, if the closest neighbor has a distance larger than the threshold, then the sample is rejected and predicted to be coming from an unknown author. Every class has its own threshold depending on the percentile, and this percentile is a hyperparameter. For example, if 0.50 is chosen, the distances above 0.50 percentile of the distribution are rejected.

3 Experiments

In the first part of the project, the baseline model was a Naive Bayes classifier; achieving F1 scores of 0.234 and 0.215 for the validation and test sets, respectively. The best performing model was Support Vector Machines (SVM), with F1 scores of 0.267 and 0.268.

The experiments in the second part are done by changing the following: feature extraction method, dimensionality of the feature space, threshold percentiles, and number of neighbors. The results of these experiments are given in Table 1.

kNN with k = 3, 5, and 7 are tried. In general, performance of 7-NN was better in terms of both F1 accuracy and rejection accuracy. Therefore, the results in Table 1 are given for 7-NN.

Table 1: Performance for open set classification

n-gram	**# features**	**Threshold**	**Rejection accuracy**	**F1* val**	**F1* test**
char 4-gram	5000	0.25	**0.659**	0.199	0.189
char 4-gram	5000	0.33	0.336	0.199	0.189
char 4-gram	5000	0.66	0.009	0.199	0.189
char 4-gram	1000	0.33	0.296	0.158	0.163
char 4-gram	1000	0.66	0.008	0.158	0.163
char 4-gram + word 1-gram	5000 + 2500	0.25	0.615	**0.207**	**0.204**
char 4-gram + word 1-gram	5000 + 2500	0.33	0.295	**0.207**	**0.204**
char 4-gram + word 1-gram	5000 + 2500	0.66	0.012	**0.207**	**0.204**
char 4-gram + word 1-gram	2500 + 2500	0.33	0.298	0.198	0.185
char 4-gram + word 1-gram	2500 + 2500	0.66	0.017	0.198	0.185

*F1 scores in this table are calculated for the closed set approach, meaning that the authors in the validation and test sets are the 50 authors in the training set.

For the closed set approach, F1 score is the highest when 5000 character level 4-gram features are used together with 2500 word level 1-gram features. As the threshold decreases, less samples meet the acception criterion. Therefore, more samples are rejected and the rejection accuracy becomes higher. On the other hand, as this threshold is smaller, there is a risk that even the samples coming from the same class are also missed and rejected.

4 Conclusion

In this project, we classified each tweet where classes represented the author, while rejecting the examples that did not belong to any author in the training set. After our experiments, we observed that 4-gram character level features, together with unigram word level features performed the best. In our initial experiments, we used various classification models. However, our open set experiments used a KNN model together with a distance threshold for rejecting examples. In our best experiment, we achieved an F1 score of 0.204, while the rejection accuracy was 0.615.

However, the difficulty of the authorship attribution task largely depends on the dataset. Another dataset with a different sampling method may produce varying results. Some reasons for that are the languages that were used, number of tweets per user, the number of users in the dataset and the time period of the dataset. We also filtered out easy users who tweeted almost identical tweets. For these reasons, comparisons to other studies is a difficult task.

Future work of this project could involve experimenting with other models, such as an SVM variant that allows us to work on open sets. We could also experiment on different feature extraction methods, such as topic modeling, sentiment analysis or deep learning methods. In addition, we could also experiment on a secondary model that accepts/rejects the nearest cluster based on the feature differences of the tweet from the centroid of the nearest cluster.

An important goal of the authorship attribution is the detection of malicious behaviors involving a single user controlling many accounts to create a large synthetic social network, or a user imitating the behavior of a reputable source. An example for that is in the recent months, some twitter users maliciously imitated the tweets of a number of corporations, causing large losses in their stock values. While the attribution is difficult for users with fewer tweets, a high quality authorship attribution model can protect such attacks on information sources and users with a large number of tweets.

References

[1] G. B. Schaalje and P. J. Fields, "Open-set nearest shrunken centroid classification," *Communications in Statistics - Theory and Methods*, vol. 41, no. 4, pp. 638–652, 2012. [Online]. Available: https://doi.org/10.1080/03610926.2010.529529

[2] A. Rocha, W. J. Scheirer, C. W. Forstall, T. Cavalcante, A. Theophilo, B. Shen, A. R. B. Carvalho, and E. Stamatatos, "Authorship attribution for social media forensics," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 1, pp. 5–33, 2017.

[3] A. Go, R. Bhayani, and L. Huang. (2009) Twitter sentiment classification using distant supervision. [Accessed 8 October 2022]. [Online]. Available: http://help.sentiment140.com/home

[4] M. A. Boukhaled and J.-G. Ganascia, "8 - stylistic features based on sequential rule mining for authorship attribution," in *Cognitive Approach to Natural Language Processing*, B. Sharp, F. SÃ¨des, and W. Lubaszewski, Eds. Elsevier, 2017, pp. 159–175. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9781785482533500081

[5] D. Jurafsky and J. H. Martin, *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, second edition.* Pearson-Prentice Hall, 2009.