Authorship Identification of Movie Reviews

CS 229 Machine Learning: Final Report

Neeraj Pradhan, Rachana Gogate, Raghav Ramesh

December 14, 2012

1. Abstract

Authorship attribution is the task of identifying the writer of a linguistic sample. The project's aim is to analyze machine learning techniques to identify authors of writing samples based on stylistic markers. Film reviews by users on the IMDb website provides an extensive dataset for authorship analysis. We use supervised learning to address the problem of identifying the author of an unknown review from a set of authors. The performance of different feature representations, specifically bag-of-words and syntactic-lexical features, is analyzed. We propose a hybrid feature representation using a compact set of syntactic-lexical features and character n-grams that gives high accuracy on the supervised learning problem while also being less computationally expensive and context agnostic.

2. Introduction

Authorship identification has historically been used in determining authorship of unattributed texts and resolving claims of plagiarism, such as the disputed authorship of the Federalist Papers as studied by Mosteller and Wallace [1]. With the proliferation of electronic text, this area is becoming increasingly important in various scenarios such as identifying email spammers, online fraud and terror message origination, trollers on message boards, and writers of malicious code.

Some past work on authorship identification has focused on email datasets or journal articles. [4] Writing samples from such datasets may not be truly indicative of a person's writing style due to external factors. For our authorship analysis study, we used movie reviews posted by users on IMDb website. The underlying hypothesis is that movie reviews would better capture individual stylistic nuances as these are sufficiently informal, uninfluenced and free from context-specific jargon.

Authorship analysis can be framed either as a supervised learning problem that requires a priori knowledge of the training class labels or as an unsupervised learning approach that uses techniques to cluster unlabeled writing samples based on similarity of stylistic markers. Some past studies on authorship identification have been limited to large text corpora by a few authors. [2-3] We try to address the problem of authorship identification in the online domain, where feature representation needs to be compact and scalable over large number of potential authors with limited writing samples.

3. Methodology and Results

We primarily address the question: 'Can we determine the author of an unclassified text from a set of *n* authors whose writing samples are available?' Extension to unsupervised approach to the problem is also discussed towards the end.

3.1 Dataset and Feature Generation

IMDb has many active users, with a good number of contributors who have reviews in excess of 200. We scraped the reviews from the IMDb website for 30 highly prolific contributors (refer to the user as in this example http://www.imdb.com/user/ur13134536/comments). Any explicit information that may aid in the classification, but is unrelated to the writing style is removed. For instance, the username of the contributor or the date on which the review was written. The average length of a movie review in our dataset was around 500 words.

These writing samples are then parsed to generate relevant features. For a preliminary analysis, we generated a bag-of-wards representation for the reviews. This gave rise to tokens (features) that are in excess of 25,000. To arrive at a more compact feature representation, we looked at stylistic features for authorship analysis. These are features which are invariant to texts written by a particular author but differ across the writings of different authors. A lot of research has been done on stylistic markers, but there is little consensus on which features constitutes a set of strong stylistic markers that can be used to differentiate authors. Juola argues that syntactic features such as function words and part of speech based features are more effective than vocabulary based features. [5]

We have chosen a set of 128 markers that is a combination of lexical, syntactic and structural features. We used python with the NLTK library for feature extraction. Some of the features used are shown below-

- Lexical – Uppercase and special characters, number of unique words, word length frequency
- Structural – Average sentence or paragraph length
- Syntactic – Occurrence of function words, or punctuations : ? . , “” : ;

3.2 Performance Analysis

We found through experimentation that a regularized logistic regression based optimizer performed best on the problem, often giving better results than SVM or Naïve Bayes. For the 10 author classification task using syntactic-lexical features, SVM gave an overall accuracy of 85% on the test data, as compared to ~90% achieved by regularized logistic regression. All our results are, therefore, with respect to an L2 regularized logistic regression using one-vs.-the rest implementation for multi-class classification. [6]

Fig. 1 shows the performance of the bag-of-words representation with that of syntactic-lexical feature based approach on the test data, as the training set size is varied. As can be seen, the latter performs better when the training size is limited. While the bag-of-words representation gives higher accuracy with more training data, the syntactic-lexical features (128 in total) achieve comparable performance. We further analyzed the top features obtained by either representation using Random Forests (RF) with 250 trees to measure variable feature importance. [7] As can be seen from Fig. 2(a), the most important features under bag-of-words approach were “2010”, “2011”, “spoilers”, etc. Moreover, bag-of-words also uses proper nouns such as movie name and director’s name. These features are highly context-specific to the domain of movie reviews.

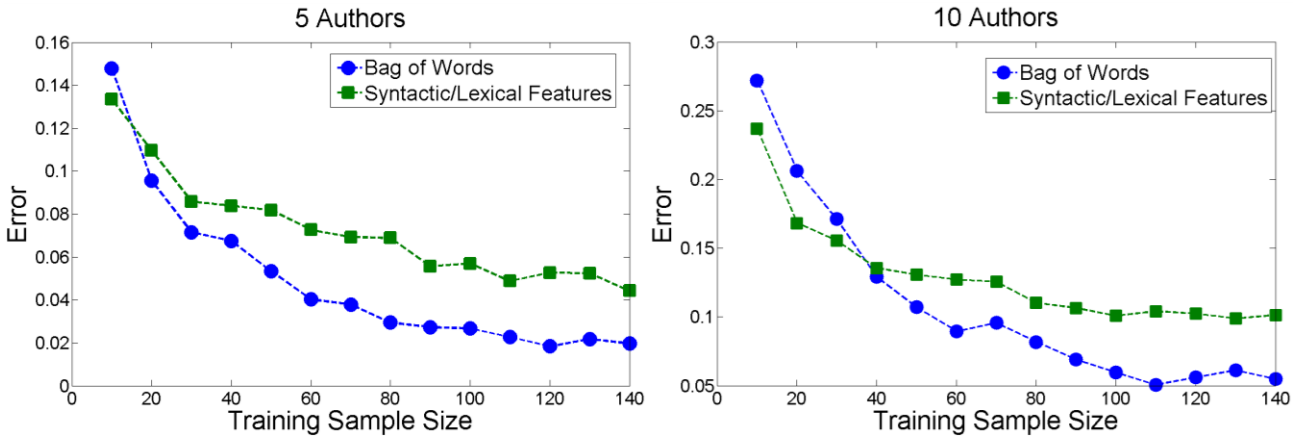

Fig. 1: *Error obtained by bag-of-words and syntactic-lexical feature representations as training sample size is varied*

Rank	Token List	Feature Importance
1	evaluation	0.0334
2	contain	0.0297
3	**spoilers**	0.0262
4	review	0.0218
5	ted	0.0171
6	film	0.0115
7	**2011**	0.0097
8	movie	0.0077
9	**2010**	0.0072
10	just	0.0069
11	really	0.0058
12	like	0.0056
13	streaming	0.0054
14	**2012**	0.0047
15	kind	0.0045

Bag-of-words

Rank	Token List	Cumulative Accuracy	Feature Importance
1	Character count	0.0743	0.0077
2	Special Characters	0.1739	0.0263
3	Space count	0.2783	0.0157
4	Lower case characters	0.3325	0.0163
5	Average sentence length (Words)	0.3833	0.0105
6	/ count	0.4281	0.0136
7	"" count	0.4688	0.0184
8	Comma count	0.5112	0.0203
9	Vocabulary count	0.5407	0.0092
10	Paragraph count	0.5849	0.0351
11	Apostrophe count	0.6156	0.0138
12	Average Word length	0.6445	0.0163
13	"I" count	0.6686	0.0089
14	Sentence count	0.6887	0.0144
15	Hapax Legomena	0.7028	0.0081

Syntactic-lexical features

Fig.2 (a), (b): *RF Feature Importance for bag-of-words and syntactic-lexical representation*

For a generalized authorship detection task, such heavy dependence on the training data may not be desirable. Thus, we consider the syntactic-lexical approach to be superior since in this case, the trained model is likely to suffer from less variance when fed with test samples that are not from the same source. For instance, when comparing emails or product reviews by the same set of writers, the heterogeneity of the context will not affect performance of the model. Given these advantages, we further probed into syntactic-lexical features to identify the most important amongst them and to prevent over-fitting. We used forward-search Feature Selection to identify the top contributing features and the incremental increase in accuracy upon each feature addition. Fig. 2(b) shows the top results from feature selection on a set of 27 authors, along with RF based feature importance. Since there is correlation in the features, we found that top 70 features contributed maximally to the classification accuracy, and the rest did not improve it appreciably.

Also, top 70 features identified by the above Feature Selection method on different sets of authors were local to that particular set of authors. Thus, a global set of features that will offer improved accuracy could not be achieved through this. Therefore, we aimed at identifying a compact global feature list that would perform well across any set of authors chosen for the supervised learning problem.

3.3 Hybrid Representation

Character n-grams have been widely used in text classification, particularly in the online medium, as they are more robust to spelling inaccuracies. [8] They can also capture more subtle stylistic nuances like the presence of punctuations (e.g. “, ”), and have therefore been previously used for authorship identification. [9] We generated a new feature set of character n-grams (*n* ranging from 2-5) and selected the top 1000 based on RF variable importance as used earlier. Using only these character n-grams as the feature set gave classification accuracies similar to that obtained by the syntactic features.

To generate a universal feature set for our problem, we used the RF based feature importance to select the top 300 features combining both the syntactic-lexical and character n-grams. We found that beyond 300, any incremental increase in features led to inconsequential improvement in accuracy. The hybrid feature set concisely captures the best features from both character n-grams and syntactic-lexical features (thus, prevents over-fitting) and remains globally applicable across different author sets. There are 85 syntactic features in this list of 300 global features selected through Feature Importance.

This hybrid feature set performed substantially better in classification of different number of authors, improving the accuracy by 40-50% over the corresponding classification using either only syntactic feature or character n-grams. Moreover, the results obtained were within ~2% of the corresponding accuracy from bag-of-words representation. This is noteworthy given that the hybrid list is compact and relatively context-free. Fig. 3 shows the accuracy obtained by the different feature representations on the authorship identification task.

Fig.4 shows the performance of hybrid feature representation relative to the syntactic lexical feature representation for classification of 5 and 10 authors as the training size is varied. Fig.4 is to be viewed in comparison with Fig.2 to better understand the improvement in accuracy resulted by the hybrid feature representation. Importantly, hybrid feature representations perform classification with high accuracy even at lower training sizes.

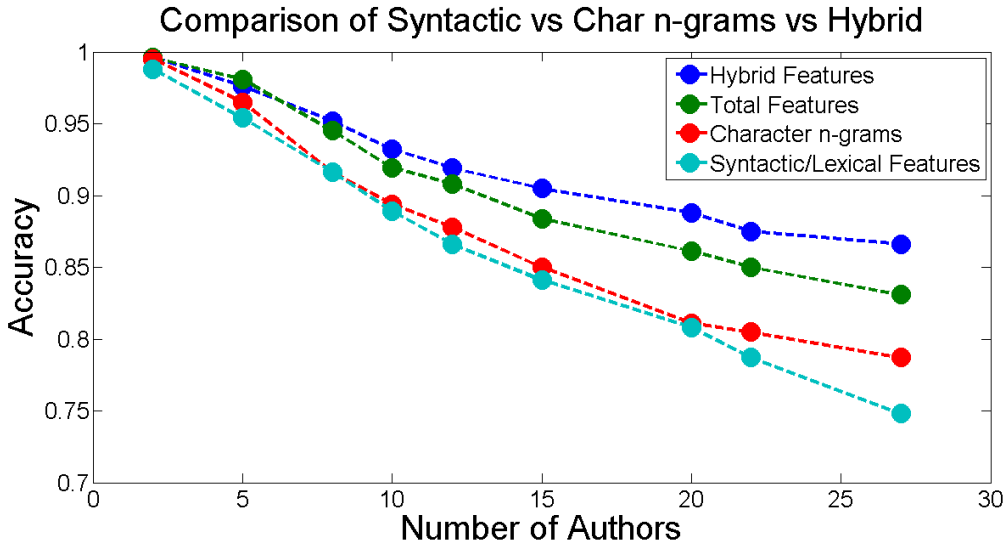

Fig. 3: Accuracy of different feature representations on the supervised learning task as the number of authors is varied.

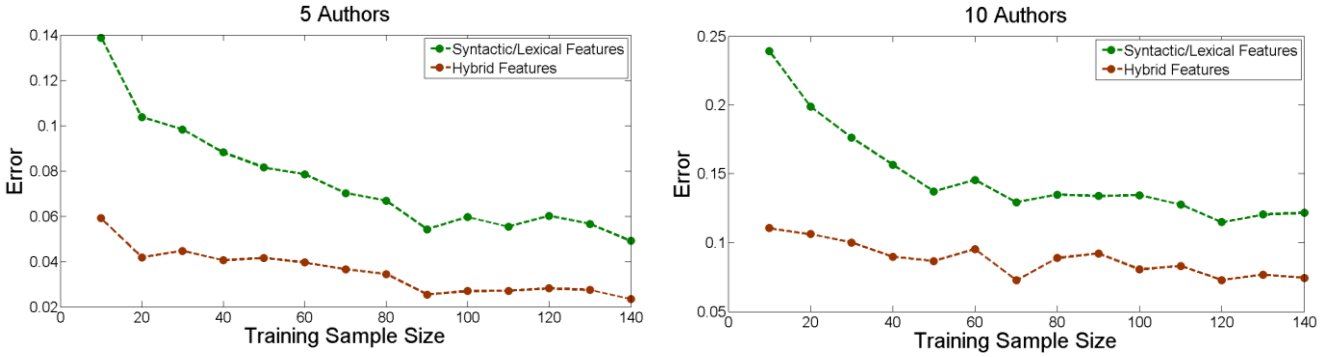

Fig. 4: *Error obtained by syntactic-lexical feature and hybrid feature representation as training sample size is varied*

4. Conclusion and Future Work

We implemented a hybrid feature representation from a compact set of syntactic-lexical features and character n-grams. This feature representation is computationally effective and context invariant, thereby providing a scalable method for author identification.

For the writing classification task when posed as an unsupervised problem, we implemented K-means clustering on the syntactic feature matrix to obtain the matching matrix. The input dataset had a total 3964 training samples for 27 authors with 128 syntactic-lexical features for each sample. The highest accuracy that we could obtain was correct classification of 99 samples out of 147 for only a single author. K-means technique generates clusters based on distance or similarity between samples, where samples within a cluster are closer to the cluster centroid. However, for high dimensional data (syntactic feature matrix has over 100 features), the distance measure may tend to clump together for different samples. An alternative approach that can be implemented, as future work, is a "grid-based" clustering methodology. [10] This methodology involves density estimation and selection of high-density areas for clustering and low density regions are assumed to be noise.

As seen earlier, the IMDb dataset that we used had a highly contextual text content that was exploited by bag-of-words to achieve high accuracy on the supervised identification task. The hybrid representation that we propose is designed to be relatively context-free. Therefore, another area for future work is to validate the results of our hybrid feature representation against a heterogeneous data source that shall test robustness against context insensitivity.

5. References

[1] F. Mosteller and D. L. Wallace. Inference and Disputed Authorship: The Federalist Series in behavioral science: Quantitative methods edition. Addison-Wesley, Massachusetts, 1964.

[2] T. V. N. Merriam and R. A. J. Matthews, "Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe," Literary and Linguistic Computing, vol. 9, no. 1, pp. 1–6, 1994

[3] Stańczyk, Ursula et al, "Machine learning approach to authorship attribution of literary texts", International Journal of Applied Mathematics and Informatics, 2007

[4] Corbin, Michael, "Authorship Attribution in the Enron Email Corpus", 2010

[5] Juola, P., "Cross-entropy and linguistic typology," in Proceedings of New Methods in Language Processing and Computational Natural Language Learning, (D. M. W. Powers, ed.), Sydney, Australia: ACL, 1998

[6] Allwein, E. L., Schapire, R. E., and Singer, Y. (2000). Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of Machine Learning Research, 1:113–141.

[7] Breiman, Leo, "Random Forests", Machine Learning, 45, 5–32, 2001

[8] William B. Cavnar and John M. Trenkle, "N-Gram-Based Text Categorization"

[9] Stamatos, E., "Ensemble-based Author Identification Using Character N-grams", In0020Proc. of the 3rd Int. Workshop on Text based Information Retrieval, pp. 41-46

[10] Michael Steinbach, Levent Ertöz, and Vipin Kumar, "The Challenges of Clustering High Dimensional Data"