Author Identification using Random Forest and Sequential Minimal Optimization

John Jenkins
Computer Science
NC A&T SU
Greensboro, USA
803-983-7195
jmjenki1@aggies.ncat.edu

Joel Bloch
Computer Science
UNC Wilmington
Wilmington, USA
919-280-6816
jvb1283@uncw.edu

ABSTRACT
Author identification is a significant factor in the global economic loss due to computer-related crimes. According to the Center for Strategic and International Studies (CSIS), an estimated 375 to 575 billion dollars is lost each year due to computer or cyber-crimes. Recently, various techniques have been used to improve the accuracy of author identification. In this paper, we propose combining unigram features and a variety of stylometric features that include n-grams and part-of-speech. With a Reuters Corpus dataset of 2,500 unique articles (50 authors with 50 news articles each), we were able to effectively capture a non-topic sensitive sample. Results with the Weka machine learning software produced classification accuracies ranging from 76.08% to 84.88% using classification techniques such as Random Forest (RF) and Sequential Minimal Optimization (SMO). Weka also ranked and weighted the most influential feature attributes. ***Dr. Kaushik Roy (kroy@ncat.edu) is the faculty advisor for this paper.***

CCS Concepts
• Security and privacy~Biometrics
• Computing methodologies~Supervised learning by classification

Keywords
Keywords—Author identification; feature reduction; classification;

1. INTRODUCTION
Biometrics is the field of research devoted to identification using physiological and behavioral characteristics [1]. Behavioral biometrics is a subgroup of biometrics based on the behavioral traits of individuals. Like physiological biometrics such as fingerprints and iris patterns, behavioral biometrics is used to find unique and quantifiable patterns. Author identification is the use of such patterns on an anonymous text to distinguish the actual authors of the text from other contestants. This task is done on a predefined set of authors and author samples [2].

Conference'10, Month 1–2, 2010, City, State, Country.

DOI: http://dx.doi.org/10.1145/12345.67890

Recently, various techniques have been used to improve the accuracy of author identification. Researchers applied unigram features such as character counts and frequencies along with the use of structural and stylometric features. Advanced features such as, n-grams (word length), function words, and parts-of-speech (POS) tags are also extracted [1][3].

2. FEATURE EXTRACTION
Unigram features signify the frequency of individual letters and characters within a text [1][4]. Previous research in author identification has shown unigram features to be some of the most accurate and efficient identifying features of an author [4]. Stylometric features include character and word counts, average characters per word, average words per sentence, and sentence count [1][2][3][4]. The stylometric features are mainly n-grams and the rate and ratio of different word lengths [2]. Similar to unigram feature extraction, ratios are collected by summing the frequency of each n-gram and dividing by the total number of n-grams. Sentiment analysis, the polarity of a word, was a key added feature as well [5]. The Stanford Parser is used to analyze the grammatical structure of sentences. Each word is parsed and given a parts-of-speech (POS) tag. The frequency of each POS tag is added to the feature vector. The ratio of each POS tag per sentence is also calculated and added. Additional features include the frequency of conjunctive adverbial phrases and contraction usage, to complete the feature vector.

3. RESULTS
Features were extracted from a Reuters Corpus dataset of 2,500 unique articles (50 authors with 50 news articles each) [6]. The selected feature vectors were extracted and placed in a CSV file used by Weka 3.6 to perform classification and cross-validation [7]. Weka is a free suite of machine learning software used to classify, analyze, and visualize data. The classifiers used were random forest (RF) and sequential minimal optimization (SMO) with a linear kernel [7][8].

Random forest is an ensemble classifier that uses decision trees, selecting nodes by a randomized procedure [8][9]. The objective of RF is to condense the variance; outputting the mode of the decisions at each tree. Sequential minimal optimization uses heuristics to split the training problem into more manageable data. SMO breaks this problem into a series of smallest possible sub-problems, which are then solved analytically. As a classifier, it is comparable to the support vector machine (SVM), which is also used for classification and regression analysis [8].

Classification and cross-validation indicates how well the features will perform in a potential identification process. The classification is done by training the program with a sample of the data set and testing the program against the remaining sample of the data. Weka's feature ranker tool was used to determine the strongest individual features and list them by weight of influence. The classification results obtained using random forest reached an accuracy of 76.08% with a cross-validation of 10 folds. Accuracy increased slightly with 76.68% for a cross-validation of 20 folds. However, for unigram features only, an accuracy of 78.88% was achieved for 10 folds. These results have us leaning toward the conclusion that the unigram features alone may be stronger in author classification.

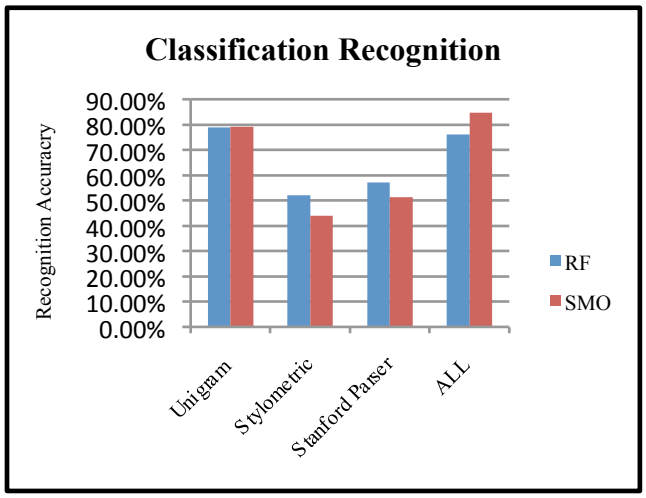

Figure 1. Classification Recognition Accuracies

The results yielded by the SMO classifier reached an accuracy of 84.68% with a cross-validation of 10 folds, increasing to 84.88% with 20 fold, for the cumulative feature vector. For unigram features alone, an accuracy of 79.08% was achieved with the cross-validation 10 folds.

Table 1. Optimal Feature Weights

Feature	Weight	Feature	Weight
E	0.636	periodFreq	0.363
EightDigit	0.473	i	0.358
CD_Freq	0.432	FiveDigit	0.357
K	0.424	LessFoutPerSent	0.354
Y	0.417	FourDigit	0.353
DigitCount	0.413	OneDigit	0.35
P	0.406	S	0.336
forwardSlash	0.405	SevenDigit	0.334
MoreSevenFreq	0.392	CD_PERCENTSent	0.331
I	0.383	CD_Count	0.329
TwoDigit	0.377	ThreeDigit	0.323
U	0.375	L	0.323
		MoreSevenPerSent	0.323

The Weka Ranker was used to identify important features (see Table 1). The ranked features encompassed mostly unigram features with the stylometric features, such as POS tags and *n*-grams, only encompassing roughly 20%. The random forest classifier was applied to the ranked features (top 62 features) and produced a classification accuracy of 75% with 10 folds. Reducing the features down to the top 25 features, the RF classification accuracy dropped to 69%.

4. Conclusion & Future Work

Unigram features have been shown in this research and previous work to produce strong results without accompanying features. Accompanying features such as *n*-grams, sentiment analysis, and the Stanford Parser parts-of-speech features have shown to produce similar results in conjunction with unigram features. The highest classification accuracy found was 84.88%, produced from the complete feature vector with sequential minimal optimization (SMO) with a linear kernel. In the future, we wish to explore new features with an objective of better accuracy, using a game-theoretic algorithm to help better determine stronger features. Future work will also include implementing program classes to test the identification accuracies produced from features and functions to find optimal weights.

5. ACKNOWLEDGMENT

This research is based upon work supported by the U.S. Government, including the NSF REU grant (#1460864) and the ARO (Contract No. W911NF-15-1-0524). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

6. REFERENCES

[1] H. Williams J. Carter, W. Campbell, K. Roy, and G. Dozier, "Genetic and evolutionary feature selection for author identification of HTML associated with malware," International Journal of Machine Learning and Computing, vol. 4, no. 3, pp. 250-255, June 2014.

[2] J. Houvardas, and E. Stamatatos, "N-Gram feature selection for authorship identification", Proc. in 12th International Conference on Artificial Intelligence: Methodology, Systems, and Applications, vol. 4183, pp. 77-86, September 2006.

[3] A. Narayanan, H. Paskov, N. Gong, John Bethencourt, E. Stefanov, E. Shin, and D. Song, "On the feasibility of internet-scale author identification," Proc. IEEE Symposium on Security and Privacy, pp. 300-314, February 2012.

[4] O. de Vel, A. Anderson, M. Corney, and G. Mohay, "Mining e-mail content for author identification forensics," ACM SIGMOD Record, vol. 30, no. 4, pp. 55-64, December 2001.

[5] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," Proc. in the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 347-354, October 2005.

[6] Reuters dataset.Available: http://about.reuters.com/researchandstandards/corpus/

[7] WEKA Classifier. Available: http://weka.sourceforge.net/doc.dev/weka/classifiers/Classifier.html

[8] J. Platt, "Fast training of support vector machines using sequential minimal optimization", Microsoft Research. Available: http://research.microsoft.com/pubs/68391/smo-book.pdf.

[9] Leo Breiman, "Random Forests" Machine Learning, vol. 45, no. 1, pp. 5-32, October 2001.