Feeling may separate Two Authors: Incorporating Sentiment in Authorship Identification Task

Braja Gopal Patra[†], Somnath Banerjee[†], Dipankar Das[*] and Sivaji Bandyopadhyay[†]
[†]Dept. of Computer Science & Engineering, Jadavpur University, Kolkata, India
[*]Dept. of Computer Science & Engineering, NIT Meghalaya, India
{brajagopal.cse,s.banerjee1980,dipankar.dipnil2005}@gmail.com,
sivaji_cse_ju@yahoo.com

Abstract

The modern era has been now extremely advanced and well developed by use of the internet especially blog, social networks, online forum and email etc. are gaining immense popularity. Thus, authorship identification is being used not only in such areas but also for forensic analysis and humanities. In this paper, we have proposed a framework for authorship identification by including the sentiment words for the classification purpose with traditional stylistic and linguistic features. We have experimented on PAN'11 dataset and achieved satisfactory results in terms of macro-average and micro-average accuracies.

1 Introduction

The task of determining the authorship of an anonymous text based entirely on internal evidence, i.e., linguistic and stylistic pattern has been gained notable research interests in recent years. More recently, it has gained greater importance due to its applications in forensic analysis, humanities, and electronic commerce. Several international evaluation tracks on authorship identification have been taken place at conferences and workshops, such as CLEF, SigWiComp etc. These conferences and evaluation tracks have aimed either at improving the results of authorship identification or verifying the systems.

The simplest form of this task can be described as: examples of the writing of a number of candidate authors are given and we are asked to determine which of them authored a given anonymous text. In this straight forward form, the authorship identification task fits the standard modern paradigm of text categorization problem (Lewis and Ringuette, 1994; Sebastiani, 2002), where documents are represented as numerical vectors that capture the statistics of potentially relevant features of the text and machine learning methods are used to find classifiers that separate the documents belonging to different classes. In this simplest form, researchers evaluate this task together with other tasks, such as topic identification, language identification, genre detection, etc. (Benedetto et al., 2002; Teahan and Harper, 2003; Peng et al., 2004; Marton et al., 2005; Zhang and Lee, 2006).

In the present work, we have tested our system on the PAN'11Dataset. We used stylistic and linguistic features and also included the frequency of the positive and negative words for the author identification task. We have seen notable improvement in terms of macro-average and micro-average accuracies while we included the sentiment features into account.

The rest of the paper is organized in the following manner. Section 2 discusses briefly the researches available till date. Section 3 provides an overview on the experimental data whereas Section 4 describes the feature selection and implementation. Section 5 presents the experiments with detail analysis. Finally, conclusions are drawn and future directions are presented in Section 6.

2 Related Work

Over the last century, varieties of methods have been applied to authorship identification tasks and can be divided into three classes of approach: a) unitary invariant approach, b) multivariate analysis approach, and c) machine learning approach. In unitary invariant approach, a single numeric function of a text is sought to discriminate between authors (Mendenhall, 1887). In the multivariate analysis approach, statistical multivariate discriminant analysis is applied to word frequencies and related numerical features (Kukushkina, 2001). In case of recent machine learning approaches, the modern machine learning methods are applied to sets of training documents to construct classifiers that can be applied to new anonymous documents (Graham et al., 2005; Zheng et al., 2006; Argamon et al., 2007).

From a machine learning point-of-view, authorship identification task can be viewed as a multi-class single-label text categorization task (Sebastiani, 2002). Several studies have been carried out based on text categorization methodologies for authorship identification or verification (Khmelev and Teahan, 2003; Peng et al., 2004; Marton, et al., 2005; Zhang and Lee, 2006). Writing style or style markers could be considered important features for classifying authorships. On the other hand, lexical, semantic, syntactic and application specific features have been used as a *stylometric* features (Holmes, 1994; Stamatatos et al., 2000; Zheng et al., 2006). Along with the previously used *stylometric* features, the feeling of the author, i.e., sentiment can also be used to identify an author. But, to the best of our knowledge no work has been carried out with this idea. So, we have experimented and found that inclusion of sentiment notably help in the task of authorship identification.

3 Corpus

In the present work, we have used the PAN'11 [1] Authorship Identification corpus for training and evaluating our author identification task. The said corpus is based on *Enron email corpus*[2] and it contains five separate training collections and seven test collections. Two training sets are provided for authorship attribution, a “Large” set containing 9337 documents provided by 72 different authors and a “Small” set containing 3001 documents provided by 26 different authors (the author sets are disjoint). The other three training sets are for verification (i.e., Verify1, Verify2 and Verify3 sets), and so these contain only emails from a single author (different from those in other training sets). The verification training sets (i.e., Verify1, Verify2 and Verify3 sets) contain 42, 55, and 47 documents, respectively.

Set ID	No of Authors	No of Documents
Large	72	9337
Small	26	72
Verify1	1	42
Verify2	1	55
Verify3	1	47

Table1: Training corpus statistics

Test corpus consists of seven sets containing a total of 286 authors with 4156 documents.

Set ID	No of Authors	No of Documents
Set-1	66	1298
Set-2	86	1440
Set-3	23	518
Set-4	43	601
Set-5	24	104
Set-6	21	95
Set-7	23	100

Table2: Test corpus statistics

Each of the training and test files is stored in an XML format, with similar schemas, as follows.

The training files look like:

```
<training>
<text file="<some unique filename>">
<author id="<unique author ID>"/>
<body>
                    TEXT OF THE MESSAGE
</body>
</text>
 ...
</training>
```

Testing files look like:

```
<testing>
<text file="<some unique filename>">
<body>
                    TEXT OF THE MESSAGE
</body>
</text>
 ...
</testing>
```

4 Experimental Methodology

4.1 Feature selection

Feature selection plays an important role in any machine learning framework and depends upon the data set used for the experiments. Thus, we have considered different combination of features to get the best results in the classification

[1] http://pan.webis.de/

[2] http://www.cs.cmu.edu/ enron/

task. The features used in this work can be categorized into three types, namely lexical, syntactic and sematic features. We have not considered the character and application feature as used in (Stamatatos, 2009). Initially, we have experimented with the traditional features into aforesaid three categories. Later, sentiment has been included as semantic features.

Lexical Feature: The natural way to view the text is as a sequence of tokens grouped into sentences. Each token corresponds to a word or a number or punctuation. Lexical features are the simple token based features like word n-gram, word frequency and sentence length etc.

Syntactic Feature: The best way to identify authorship is to implement syntactic information. The main idea is that authors tend to use some similar syntactic pattern unconsciously (Stamatatos, 2009). Therefore, syntactic feature is more important than the lexical feature and the use of syntactic feature can be seen in (Houvardas and Stamatatos, 2006). The syntactic feature includes the sentence and phase structure, errors in writing, part of speech and chunks etc.

Semantic Feature: It is clear that the more detailed text analysis is required for extracting the stylometric features. More complicated tasks such as full syntactic parsing, semantic analysis, or pragmatic analysis cannot yet be handled adequately by current NLP technology for unrestricted text (Stamatatos, 2009). As a result, very few attempts have been made to exploit high-level features for stylometric purposes. The semantic features include synonyms, semantic dependencies, positive and negative word frequencies etc.

Initially, we experimented with the features like simple Unigram, Bigram and Trigram (Houvardas and Stamatatos; 2006). Then we have included all stylistic features like number of stop words, list of foreign words, list of punctuations and list of pronouns etc. Again, we have included the frequency of positive and negative word class frequencies.

Unigram: All the tokens, which have not been marked as stop words, punctuations and foreign words, are listed in the unigram list. We have kept a threshold frequency for discarding all the lower level unigrams. In our experiments, we have considered unigrams occurred more than 500 times.

Bigram and Trigram Frequency: Bigrams and trigrams are common features for author identification (Houvardas and Stamatatos; 2006). It is found that the authors have tendency to reuse the same phrase. Thus, we have used the threshold frequency. As the 4-grams and 5-grams are important features in the task (Houvardas and Stamatatos; 2006), in contrast, we have not included them in our experiment as the corpus is small and the frequencies were also negligible.

Stop words frequency: Stop words have been found as one of the important features. A total of 329 stop words have been prepared manually.

List of Foreign Words (FW): These are thewords, which are tagged as FW by the Stanford-CoreNLP POS tagger[3]. These are basically “meee”, “yesss”, “thy”, “u” and “urs” etc.

List of Punctuations: 10 types of punctuations are prepared manually.

List of Pronouns: The frequencies of the pronouns are also computed. Pronouns are tagged as PRP by StafordCoreNLP POS tagger.

Average Length of Word and Sentence: We have considered the average word and sentence length in documents. The sentence boundary is detected by the StanfordCoreNLP tool.

Positive and Negative word class: It is found that the positive and negative word classes are also key features for automatic author identification. Thus, these two classes contain the words which are not listed in our existing unigram list. After getting all possible POS from RiTaWordNet[4], the sentiment scores of the words have been calculated using the SentiWordNet 3:0[5] lexicon. Then, the words having sentiment score greater than 0.1 and less than -0.1 (threshold value: |0.1|) have been considered as the positive and negative sentiment word classes.

It has been found that the size of each document varies, i.e., some documents contain more number of words and some documents contain less words. So, we have normalized each bag of word feature by dividing the total number of words in a document.

4.2 Experimental Setup

After removing the XML tags from the documents, various NLP tools have been applied to identify the features. The Stanford CoreNLP package has been used to detect the sentence boundary and then, the average word in a sentence has been calculated. Each word of the XML document has also been stemmed by the Stanford CoreNLP package. The punctuation list

[3] http://www-nlp.stanford.edu/software/corenlp.shtml
[4] http://www.rednoise.org/rita/wordnet/documentation/
[5] http://sentiwordnet.isti.cnr.it/

has been used to calculate frequencies from the text.

The Stanford CoreNLP POS tagger has been used to tag parts of speech (POS) of each word. Pronouns and Foreign words frequencies have been calculated from the tagged text. Pronouns and Foreign Words are tagged as PRP and FW by the Stanford CoreNLP POS tagger.

Word class frequencies have also been calculated by using the manually prepared lists. The Positive and Negative word frequencies have also been determined by using the RiTaWordNet. The stop word frequency has been determined by using stop words list. The frequencies of n-gram, positive word, negative words, pronoun, punctuation and foreign words have been normalized by dividing the total number of words present in the document. The extracted features are also used to prepare our test templates.

We have used the API of Weka 3.7.7.5[6] to accomplish our classification experiments. Weka is an open source data mining tool. It presents a collection of machine learning algorithms for data mining tasks. We employed the Decision tree (J48) for classifying the documents. The decision tree model has been trained by training template and the model has been used to classify the test template.

5 Results and Discussions

We have used the same evaluation strategy as defined in the PAN'11 author identification task (Argamon and Juola, 2011). PAN'11 used the standard information retrieval metrics of precision, recall, and F1. Precision, for a particular author A, is defined as the fraction of attributions that a system makes to A that are correct:

$$P_{\mathrm{A}} = \frac{correct(A)}{attributions(A)}$$

Recall, for a particular author A, is defined as the fraction of test documents written by A that are (correctly) attributed to A:

$$R_{\mathrm{A}} = \frac{correct(A)}{documents{-}by(A)}$$

F1 is defined as the harmonic mean of recall and precision:

$$F_1 = \frac{2\, P_A\, R_A}{P_A + R_A}$$

Two methods namely macro-averaging and micro-averaging have been applied to aggregate these measures over all the different test authors. For a given metric M, set of n authors {A$_i$}, with a total of k test documents, these are defined as:

$$\text{macro-avg}_{\mathrm{M}}(\{A_i\}) = \tfrac{1}{n}\textstyle\sum_i M_{A_i}$$

$$\text{micro-avg}_{\mathrm{M}}(\{A_i\}) = \tfrac{1}{k}\textstyle\sum_i k_i M_{A_i}$$

Where k_i is the number of test documents written by author A$_i$. Micro-averaging will give more credit to accuracy on authors with more test documents, while macro-averaging gives the same credit to all authors, even if they wrote just one test document.

To establish the effects of sentiment feature in the author identification task, each of the test set corpus have been evaluated twice- first, without sentiment feature represented by F and second, with the sentiment feature represented by F+S. The detail of experiment results have been shown in the table as follows-

	Macro-avg			Micro-Avg		
	P	**R**	**F1**	**P**	**R**	**F1**
F	54.3	53.6	53.9	62.4	58.9	60.6
F+S	67.5	63.1	66.1	72.6	64.7	68.4

Table3: Large test set without extraneous documents

	Macro-avg			Micro-Avg		
	P	**R**	**F1**	**P**	**R**	**F1**
F	72.5	56.1	63.3	78.3	57.4	66.2
F+S	83.7	63.1	72.0	86.5	64.7	74.0

Table4: Large+ test set with extraneous documents

	Macro-avg			Micro-Avg		
	P	**R**	**F1**	**P**	**R**	**F1**
F	62.3	49.1	54.9	66.7	55.1	60.3
F+S	72.9	56.3	63.5	78.2	66.1	71.6

Table5: Small test set without extraneous documents

	Macro-avg			Micro-Avg		
	P	**R**	**F1**	**P**	**R**	**F1**
F	76.1	58.6	66.2	78.2	59.0	67.3
F+S	86.9	67.1	75.7	89.1	69.7	78.2

Table6: Small+ test set with extraneous documents

[6] http://weka.wikispaces.com/Use+WEKA+in+your+Java+code

It has been observed from the experiments that inclusion of sentiment features namely positive and negative word classes improves the accuracy notably for each of the test sets. Our system without sentiment feature (i.e., without positive and negative word class) gives almost the same accuracy compare to the results of PAN'11 (Argamon and Juola, 2011). Considering sentiment feature improves the macro-average accuracy by 13.2%, 11.2%, 10.6% and 10.8% respectively and micro-average accuracy by 10.2%, 8.2%, 11.5% and 10.9%, respectively.

6 Conclusions

The main contribution of this work is successfully introducing sentiment features as semantic features in the task of authorship identification. In this work, we presented a system for author identification task performed on the PAN'11 dataset. We have included the traditional stylistic and linguistic features. We have also introduced the sentiment features like positive and negative word class frequency with traditional features to the author identification task, which results in increase of accuracy in terms of macro-average and micro-average accuracy.

In our future work, the accuracy of the classification can be improved by finding and incorporating more traditional as well as sentiment features in the task of authorship identification. It would also be interesting to perform deeper features engineering for finding demographic and psychometric author traits more correctly.

Acknowledgement

The work reported in this paper is supported by a grant from the India-Japan Cooperative Programme (DST-JST) 2009 Research project entitled “Sentiment Analysis where AI meets Psychology” funded by Department of Science and Technology (DST), Government of India.

References

Braja G. Patra, Somnath Banerjee, Dipankar Das, Tanik Saikh and Sivaji Bandyopadhyay. 2013. Automatic Author Profiling Based on Linguistic and Stylistic Features. CLEF 2013 Evaluation Labs and Workshop–Working Notes Papers.

Braja G. Patra, Amitava Kundu, Dipankar Das and Sivaji Bandyopadhyay. 2012. Classification of Interviews – A Case Study on Cancer Patients. In *Proceedings of 2nd Workshop on Sentiment Analysis where AI meets Psychology* (SAAIP-2012), pages 27-36.

Carole E. Chaski. 2001. Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, 8:1-65.

Dario Benedetto, Emanuele Caglioti and Vittorio Loreto. 2002. Language trees and zipping. *Physical Review Letters*, 88(4):0487021-0487024.

David D. Lewis and Marc Ringuette. 1994. A comparison of two learning algorithms for text categorization. In *Proceedings of the Third annual symposium on document analysis and information retrieval*, vol. 33, pages 81-93.

David Madigan, Alexander Genkin, David D. Lewis, Shlomo Argamon, Dmitriy Fradkin and Li Ye. 2005. Author identification on the large scale. In *Proceedings of the Meeting of the Classification Society of North America*.

Dell Zhang and Wee S. Lee. 2006. Extracting key-substring-group features for text classification. In *Proceedings of the 12th Annual SIGKDD International Conferenceon Knowledge Discovery and Data Mining*, pages 474-483. ACM.

Dmitry V. Khmelev and William J. Teahan. 2003. A repetition based measure for verification of text collections and for text categorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 104-110. ACM.

Efstathios Stamatatos, Nikos Fakotakis and George Kokkinakis. 2000. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4), 471–495.

Efstathios Stamatatos. 2006. Ensemble-based author identification using character n-grams. In *Proceedings of the 3rd International Workshop on Text-based Information Retrieval*, pages 41-46.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3): 538-556.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM computing surveys* (CSUR), 34(1): 1-47.

Fuchun Peng, Dale Schuurmans and Shaojun Wang. 2004. Augmenting naive Bayes classifiers with statistical language models. *Information Retrieval*, 7(1): 317-345.

Harald Baayen. 1994. Authorship attribution. *Computers and the Humanities*, 28(2):87–106.

John Houvardas and EfstathiosStamatatos. 2006. N-gram feature selection for authorship identification. In *proceedings of the Artificial Intelligence: Methodology, Systems and Applications*, pages 77-86, Springer Berlin Heidelberg.

Malcolm Coulthard. 2004. Author identification, idiolect, and linguistic uniqueness. *Applied linguistics*, 25(4): 431-447.

Neil Graham, Graeme Hirst and Bhaskara Marthi. 2005. Segmenting documents by stylistic character. *Natural Language Engineering*, 11(4): 397-415.

O. V. Kukushkina, A. A. Polikarpov and Dmitry V. Khmelev. 2001. Using literal and grammatical statistics for authorship attribution. Problems of Information Transmission. 37(2): 172-184.

Olivier D. Vel, Alison Anderson, Malcolm Corney and George Mohay. 2001. Mining E-mail content for author identification forensics. ACM Sigmod Record, 30(4): 55-64.

Patrick Juola and Efstathios Stamatatos. 2013. Overview of the Author Identification Task at PAN 2013. CLEF 2013 Evaluation Labs and Workshop–Working Notes Papers.

Rong Zheng, Jiexun Li, Hsinchun Chen and Zan Huang. 2006. A framework for authorship identification of online messages: Writing style features and classification techniques. *Journal of the American Society of Information Science and Technology*, 57(3): 378-393.

Shawndra Hill and Foster Provost. 2003. The myth of the double-blind review?: author identification using only citations. ACM SIGKDD Explorations Newsletter, 5(2): 179-184.

Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan R. Hota, Navendu Garg and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802-822.

Shlomo Argamon and Patrick Juola. 2011. Overview of the International Authorship Identification Competition at PAN-2011. *CLEF (Notebook Papers/Labs/Workshop).*

Stefano Baccianella, Andrea Esuli and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10),* Valletta, Malta, May.

Thomas C. Mendenhall. 1887. The characteristic curves of composition. *Science,* 214S: 237-246.

William J. Teahan and David J. Harper. 2003. Using compression-based language models for text categorization. In *Proceedings of the Language Modeling for Information Retrieval*, pages 141-165. Springer Netherlands.

Yuval Marton, Ning Wu and Lisa Hellerstein. 2005. On compression-based text classification. In *Proceedings of the Advances in Information Retrieval*, pages 300-314. Springer Berlin Heidelberg.