UOIuBIH
ORSinBIH
Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: http://scjournal.ius.edu.ba

IUS Soft Computing Research Group

Better Features Sets for Authorship Attribution of Short Messages

Nesibe Merve Demir

International University of Sarajevo,
Faculty of Engineering and Natural Sciences,
HrasnickaCesta 15, Ilidža 71210 Sarajevo,
Bosnia and Herzegovina
ndemir@ius.edu.ba

Article Info

Article history:
Article received on 17 Jan. 2017
Received in revised form 17 Feb.2017

Keywords:
Authorship attribution, feature sets, support vector machine, short text's classification.

Abstract

Authorship authentication analysis can help to display information about the writers of messages by analyzing the writing styles. Previous researches in the authorship authentication were showed that generally people have their unique stylistic discriminators and characteristics, just like their fingerprints or signature. In this concept, researchers are developing different analysis features and techniques and have gained remarkable results in the authorship identification research field.

Authorship authentication of online messages became an outstanding research topic in the last decades because of internet usage growth. One of the problems of authorship authentication analysis regarding online sources is short messages usage. Author identification techniques are started to be applied to short and informal texts in last decade and get very significant results.

Authorship authentication is one of the security concerns in social network and in this research we will study how to authenticate a user by the writing style in a short text posted on Twitter. The effects of different feature sets and sample sizes are evaluated in the research.

1. INTRODUCTION

The process of verification of a text's author legitimacy is called as authorship authentication in which the security interests regarding online social networks would be addressed (Li Y, 2015). State-of-the-art social network sites in current times only apply textual login and password for authenticating their users. Like an additional authentication mechanism, the written messages may be applied for the authenticating of users.

The study has been usually conducted in long text, even though authorship authentication with the application of stylometry has been a prominent field. For the identification of authors with regard to the style of their writings; linguists, computer engineers and scholars of the humanities have been together contributing to the development of automated methods during the past 50 years (Rocha, 2016). Each author has unique habits which have impacts on the forms and contents of written texts of them. With the application of machine learning algorithms, these characteristics generally can be quantified.

The study of linguistic style which is related with features like syntactic structure, word choice, word count, and sentence length is called as stylometry. With the consideration of stylometric features, stylometry expresses personal writing styles (Iqbal, 2010). Stylometric features should be found for the adaptation of authorship attribution to social media in which the diversity of the language deployed is captured. For example, a tweet only 140-character long does not produce an important amount of information at the word-level or from syntactic structure of it.

Through several related stylometric researches, different combinations of features are present on current day. According to Kacmarcik and Gamon, for the authorship attribution, the feature selection is among the most essential part of it. With the feature selection, an easy formation of word frequencies in document verification is possible for the researcher.

2. RELATED WORKS

Authorship authentication in social media messages is a challenging problem because of the short text issue. Scientists had been focusing on this problem before the social media had not yet been developed as a part of forensic stylometry in e-mails in which short form writing is the norm. Including similarity measures and SVM, some strategies perform in a direct way, although better performance can be obtained by means of features and classification approaches which are custom-tailored for the use of attribution problems together with texts' small samples.

Anderson et al. [6] and de Vel et al. (2001) focused on a variety of character-level statistics including white space, capitalization, and punctuation counts to compensate concerning a small amount of information subsistent within such a context.

Moreover, Forstall and Scheirer (2011) suggested that character-level n-grams operate by being useful proxies for phonemes which reflect the sound of words implying another facet of language which can be measured as a feature for authorship attribution.

Jenny et al. (2016) compared classifiers and features for Facebook posts, an average of 20.6 words. They had an average accuracy rate of 79.6% for 30 users with 233 features. They concluded that in their research SVM and decision trees produced comparable results and best performance.

Paulo et al. (2016) used 179 syntactic features and SVM classifier and reach 75.5% accuracy for English language short texts. They used 20 samples for each of the 20 authors. In their work they studied two models, the writer-dependent and writer-independent.

Rocha et al. (2016) wrote a review about authorship attribution for social media forensics. They collected data from Twitter and used 50 authors' tweets. After achieving low results in experiences, they tried power mean SVM with 1000 tweets and the highest accuracy rate they achieved around 70%.

FEATURES USED IN AUTHORSHIP AUTHENTICATION OF SHORT TEXTS

Researchers built feature sets of writing-styles. Based on the review of the studies, (Zheng, 2006) integrated four types of features into the feature set: lexical, syntactic, content-specific, and structural features. Lexical features can be further divided into character based and word-based features. Syntactic features, including function words, punctuation, and part of speech, can seize an author's writing style at the sentence level. The discriminating power of syntactic features is derived from people's different habits of organizing sentences. Structural features represent the way an author organizes the layout of a piece of writing. De Vel (2000) introduced several structural features specifically for e-mail. Anderson et al (2001) turned to a variety of character-level statistics such as capitalization, white space, and punctuation counts to countervail the small amount of information inherent. Content-specific features are important discriminating features for online messages. The selection of such features is dependent on specific application domains. On the Web, a writer may frequently send online messages including a relatively small range of topics whereas different users may distribute messages on different topics. For this reason, special words or characters closely related to specific topics may provide some clue about the identity of the author. New features are added for online messages' authentication because of the need of special characteristics.

Features are grouped roughly under five titles:

Lexical Features: This type has two subtypes; character-based and word-based. They are the earliest features used in the analysis and represent an author's lexicon-related writing styles. Yule (1944) employed features like sentence length and vocabulary richnesses. Later, Burrows (1992) added a set of more than 50 high-frequency words. Holmes (1998) examined shorter words (two or three letter words). Zheng et al. (2006) adopted 87 lexical features for English online messages. Some of these features are given below.

Table1. Lexical features

Character-based features	Word-based features
Total # of characters	Total # of words
Total # of upper-case characters	Total # of short words
Total # of digit-characters	Total # of characters in words
Total # of white-space characters	Average word length
Total # of tab spaces	Average sentence length in terms of character
Frequency of letters	Average sentence length in terms of word
Frequency of special characters	Total different words
	Hapax legomena*
	Hapax dislegomena*
	Yule's K measure*
	Simpson's D measure*
	Sichel's S measure*
	Brunet's W measure*
	Honore's R measure*
	Word length frequency distribution*

Note: The definitions of measures with '' can be found in Tweedie and Baayen (1998)*

Syntactic Features: This type includes punctuation and function words. They can capture an author's writing-style at the sentence level. These features created for capturing people's different habits of organizing sentences. Mosteller and Wallace (1964) first used the frequency of occurrence of 30 function words. Different sets of function words have been tested in various studies. There are not generally accepted good sets of function words. Zheng adopted a large set of 150 function words that are selected based on previous researches.

Structural Features: They represent the author's habits when organizing a piece of text. Paragraph length, use of indentation and use of signature can be strong evidence of personal writing style. De Vel et al (2000) employed firstly several structural features specifically for e-mail. Zheng adopted 14 structured features, four of them newly proposed. These features are listed in the below.

Table 2. Structural features

Total # of lines	Has separators between paragraphs
Total # of sentences	Has quoted content
Total # of paragraphs	Position of quoted content
Total # of sentences per paragraph	Indentation of paragraph
Total # of characters per paragraph	Use e-mail as signature
Total # of words per paragraph	Use telephone as signature
If has a greeting	Use url as signature

Content-specific Features: They refer to words in a specific topic and such features are depend on specific application domains. Special words which are closely related to specific topics may provide clue about the identity of the author. Zheng identify 11 key words as content specific features.

Social Network-specific Features: J. S. Li et al (2016) added six features. They claimed that these features reflect a more causal writing style. Social media writers have a style that is commonly seen in chats and they write in a vulgar way that is similar to daily conversation. Features that were included are listed in the below.

Table 3. Social network-specific features

Happy-face emotion	Sad-face emotion
Abbreviations (LOL)	Ending a sentence without a punctuation mark
Starting a sentence without upper-case letter	Not mentioning I or We in the post

Another type of features is the top character n-grams, or a sequence of n-characters. These features may be related to syntactic structure of texts. It is known also as POS tagging. Some researches have been done with this type of feature sets.

3. PROPOSED METHOD

We extended the approach followed in our previous work (Demir,2016).We used different feature sets with different number of testing groups to identify the effect of features and number of tweets per author.

For authorship attribution, a text's words are basically convenient features. Nevertheless, it is not possible to treat basically all of the words as features. The scrap of the function words is a prevalent thing that these words often exist even though they do not transmit much in the case any semantic meaning for isolating a more stable signal. But, in some instances, function words may be full of use for attribution. Essentially, statistics associated to function words still exist as input to some algorithmic approaches (M.Koppel, 2007), even though being among the earliest features proposed for manual authorship attribution (M.Mascol, 1888). It is important to choose functions words carefully.

Rudman (1998) summarized around 1000 features for authorship authentication applications in English language. Zheng et. al (2006) adopted a set of features based on previous literature for online messages including 270 features for English.

Studies showed that different types of features have different power of discrimination. Therefore it is important to identify the key features.

We collected 106 features. They contained Character-based lexical features, word-based lexical features, syntactic features, structural features and social networking-based features (see Table 4). 71 of them are function words. Function words are selected from the list that was prepared by Zheng et al. (2006).

Table 4. Features used in this research

Reference Number	Features	Description
1	Total number of characters	
2	Total number of characters with space	
3	Total number of words	
4	Average word length	
5	Total number of sentences	
6	Total number of uppercase	
7	Total number of lowercase	
8	Total number of short words	Less than four characters
9-31	Frequency of letters	A-Z
32	Frequency of emotions	Smiley face":)"; sad face ':('
33	Frequency of 'LOL'	
34	Total number of special characters	
35	Total number of punctuations	, . ! #@$?;*+-/
36-106	Frequency of function words	71 words

The choice of a classification method is the following stage after a feature set has been selected. Support vector machine was used in our research. We used competing team that would vote for input to choose the correct author (Demir, 2016). At the validation process, created competing team receives data from writers with the same amount of tweet that was used in the training set and decides which writer owns the tweets.

4. DISCUSSION

Users with distinctive writing styles are easier to be distinguished from others. We chose two users who have more distinctive writing styles than others so we can experience the effect of features. Our first aim in this research was analyzing the effect of features and showing if the number of tweet affects the results. We tried different number of tweet as input. The results showed that less number of tweets has higher accuracy (See table 5). We later used different number of features. Using all features received low results. 16 features with 20 tweets had the highest accuracy, 95%. In addition, using more features costs more computational time to calculate values. While having similar results, it is necessary to decide on the trade-offs between computational effort and accuracy. In our case, 16 features seem more beneficial.

Table 5. Testing different numbers of tweets and features for chosen two users' Twitter data

Test	# of tweet	# of features	AR	HAR
1	100	37	86	89
2	60	37	86.5	90
3	40	37	85	92
4	20	37	92.5	95
5	20	106	50	100
6	20	16	95	95
7	20	30	90	95

AR: Average accuracy rate, HAR: Highest accuracy rate

After receiving these results, we wanted to check different groups of features set. By choosing randomly 16 features, we continue to test same two users. After experience several different sets of feature, we concluded that the accuracy rate did not increase by using different feature sets (See table 6).

Table 6. Sample of testing different features on two authors (for feature number, see table 1)

# of tweet	features	AR	HAR
20	1-16	83	90
20	1-12,34-37	82	95
20	1-10,31-36	85	95
20	17-32	75	75
20	1-5,7-8,31-35,102-106	85	90

AR: Average accuracy rate, HAR: Highest accuracy rate

We create another test to see interaction between users. We tested the chosen user and the rest by two pairs (See table 7). Accuracy rate was 78.94%. The highest accuracy rate was again 95%.

Table 7. Testing a chosen user with the remaining 33 users pair wise

# of tweet	# of features	AR	HAR
20	16	78.94	95

AR: Average accuracy rate, HAR: Highest accuracy rate

When we test ten users with 16 features, the highest accuracy was 65%. This result shows that number of users also affect the accuracy.

5. CONCLUSION

The generation of optimal feature set is the challenging future work.

We experienced that features used in the current research is not enough good for achieving better results. We need to expand our feature set.

The improvement direction will be finding optimal sets of features which will improve the classification accuracy. It is not just for avoiding computationally expensive sparse feature representatives but also to choose best sets of the features. The best feature sets that can be used in the short texts' authentication problem are needed to be chosen. By choosing minimally required features, competitive classification accuracies in conjunction could be achieved with our classifiers.

REFERENCES

Anderson , A., De Vel, O. Y., Corney , M. and Mohay, G. (2001) Mining Email Content for Author Identification Forensics, Sigmod Record, vol. 30, pp. 55-64

Burrows, J.F. (1992), Word patterns and story shapes: The statistical analysis of narrative style. Literary and Linguistic Computing 2, pp. 61–67

De Vel, O. Y. (2000), Mining Email Authorship, Proceedings of Paper presented at the Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining (KDD)

Holmes, D. (1998), The evolution of stylometry in humanities scholarship, Literary and Linguistic Computing, vol. 13, no. 3, pp 111-117

Iqbal, F., Khan, L. A., et al. (2010), E-mail authorship verification for forensic investigation, Proceedings of the 2010 ACM Symposium on Applied Computing, pp. 1591-1598.

Li, J. S., Chen, L. -C., Monaco, J. V., Singh, P., and Tappert, C. C. (2016) A comparison of classifiers and features for authorship authentication of social networking messages, Concurrency and Computation: Practice And Experience, Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/cpe.3918.

Li Y, Dai W, Ming Z, Qiu M. (2015), Privacy protection for preventing data over-collection in smart city. IEEE Transactions on Computers, p.1.

Mosteller, F. and Wallace, D.L. Inference and Disputed Authorship: The Federalist. Addison-Wesley, 1964

Rocha, A. et al. (2016), "Authorship Attribution for Social Media Forensics," in IEEE Transactions on Information Forensics and Security, vol. 12, no. 1, pp. 5-33

Rudman, J. (1998), The State of Authorship Attribution Studies: Some Problems and Solutions, Computers and the Humanities, vol. 31, pp 351-365

Yule, G.U. The Statistical Study of Literary Vocabulary. Cambridge University Press, 1944

Zheng, R. Li, J., Chen, H. and 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, vol. 57, no. 3, pp. 378–393

Demir, N.M. (2016), Authorship Authentication for Twitter Messages Using Support Vector Machine, Southeast Europe Journal of Soft Computing, Vol. 5, No. 2, pp. 1-6

Tweedie, F. J., and Baayen R. H.,(1998) How Variable May a Constant Be? Measures of Lexical Richness in Perspective, Computers and the Humanities, vol. 32, no. 5, pp. 323–352.