VOL.5 NO.2 September 2016 - ISSN 2233 – 1859

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

Authorship Authentication for Twitter Messages Using Support Vector Machine

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 16 May 2016
Received in revised form Sept 1 2016

Keywords:
Authorship attribution, Artificial Intelligent, support vector machine (SVM), short texts

Abstract

With the rapid growth of internet usage, authorship authentication of online messages became challenging research topic in the last decades. In this paper, we used a team of support vector machines to authenticate 5 Twitter authors' messages. SVM is one of the commonly used and strong classification algorithms in authorship attribution problems. SVM maps the linearly non separable input data to a higher dimensional space by a hyperplane via radial base functions. Firstly using the training data, 10 hyperplanes that separate pair wise five authors training data are built. Then the expertise of these SVMs combined to classify the testing data into five classes. 20 tweets with 16 features from each author were used for evaluation. In spite of the randomly choice of the features, one of the author accuracy around 75% is achieved.

1. INTRODUCTION

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.

One of the problems of authorship authentication analysis regarding online sources is the huge quantity of online data and a big part of candidate authors which make it more difficult. The other difficulty occurs because of short messages usage in social media while working with online textual data.

Author identification techniques are also started to be applied to short and informal texts in last decade with this change and get very significant. E-mail, forums and messaging boards, blogs, social networking sites such as Facebook and text messages are among the sources.

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.

2. AUTHORSHIP ATTRIBUTION

One of the main concerns in Authorship Attribution is the search for quantifiable features that are able to differentiate between authors of some text, which can be used in literary tasks of textual analysis for works edited, translated, with disputed authorship or anonymous, but also with forensic aspect in view to detect plagiarism, forgery of the whole document or its constituent parts, verify ransom notes, etc.

Author identification analysts claim that each writer possesses some unique characteristic, called the authorial or writer invariant that keeps constant for all texts written by this author and perceivably different for texts of other authors. To find writer invariants there are used style markers which are based on textual properties belonging to either of four categories: lexical, syntactic, structural, and content-specific (Cyran, 2007).

A wide range of classification algorithms was used to identify authors. They include component coupling, neural network, Bayesian classifier, logistic regression, decision tree, covariant or linear discriminant algorithm, principal component analysis, nearest neighbor, rough sets and support vector machine (SVM).

3. STYLOMETRY

Regarding to a stylometric study, the most reliable data is considered as an author's linguistic style which has particular features without dependence on author's will because it is not possible for the author to manipulate the features as knowingly. Features of the author's style are aimed to describe by stylometry as well as identification of the statistical methods in order to know the similarity which occurs between two or more textual sources and the features.

A new age is characterized with the presence of stylometry. During the last decades, for the identification of authorship, a wide range of mathematical tools are developed including statistical tests and Artificial Intelligence Techniques. The tools have used by scientists for texts including large spectrum of literary genres and time periods. The Federalist Papers; Civil War Letters; Shakespeare's plays; The New Testament; The Royal Book of Oz, and The Dialogues of Plato are among them (J. Binongo, 1999, E. Charniak, 1993).

Stylometry application for the identification of authorship extends to the time of precomputer. The application of stylometric features to textual analysis took attention of many researchers that Mendenhall as an American physicist offered in the end of 1880s authorial styles may be 'fingerprinted' with the act of counting the numbers of letters in the words that they used. In the beginning of 1990s, Yule developed the use of counting the features of a text by covering the lengths of sentences in the text. In order to test Greek prose authorship, an application of sentence-lengths was performed by Morton, in the middle 1960s. Individuals as "human computers" were used by Zipf, a Harvard University German Professor for counting how much time each word is seem in a text and therefore, for ranking the frequency of certain words.

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. They have the limited work for word frequencies since these features are usually recognized as the good way for the identification of authorship attribution. With the feature selection, an easy formation of word frequencies in document verification is possible for the researcher. The discriminatory potential gets higher in the case that more than one feature is applied as a combination with each other.

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 capture 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. Content-specific features are important discriminating features for online messages. The selection of such features is dependent on specific application domains. On the Web, one user may often post online messages involving 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.

4. HISTORICAL BACKGROUND

If we check authorship identification on online texts, most of the studies were done on email and blog posts. They are relatively short texts if we compare with literary texts. Also the researches are done on online texts have experimented different situations like having many authors or many posts by an author.

De Vel et al. (2001) worked on two kinds of features for assisting in determination of the author of the email in their studies in which application of authorship authentication was used on a set of 150 emails which are gathered from three authors. Firstly, style marker features were taken from like application of quotations, the number of spaces used in the emails or uppercase letters, and so on additionally to another type of features that were structural features such as salutations at the end of emails. In order to conduct experiments, the Support Vector Machine (SVM) was used in the work.

Another work conducted by Zheng et al. (2003) which extended the work of De Veletal. with application of authorship authentication on web forum messages, not emails. English and Chinese languages were the two languages of messages in which SVM, Artificial Neural Network (ANN) and decision tree classifiers were applied in their experiments.

Clark and Hannon (2007) had their study on authorship authentication. They worked on the choices of the author for synonyms in the writing. The authors argued that it is sufficient to know favored kind of synonym that an author uses for discovering the identity.

Linear SVMs for text categorization was utilized by Dumais et al. The reason for the application is their accuracy and fastness. They appear as 35 times faster for training as a comparison to the next most accurate (a decision tree) of the tested classifiers. SVMs were utilized within the Reuter-21578 collection, emails and web pages

by them. Emails were separated as spam and non spam by Drucker et al. According to their findings, boosting trees and SVMs have similar performance with respect to accuracy and speed of them. The thing is that the training of SVMs is so faster in a significant way.

Rachid et al. (2009) studied for a framework of email forensic analysis. They considered 63 e-mails from 3 senders. Support vector machine and C4.5 (decision tree) were used in the experiments. Average accuracies were between 69% and 83%. They have proposed a new technique of mining style variation to capture the style changes of authors.

Narayanan et al. (2012) assembled a dataset covers almost billion words from 100000 blogs. They tried several classifiers and features. Depend of the model their success is from 20% to 80%.

Brocardo et al. (2014) tried a supervised learning technique combined with n-gram feature set for short email messages. They used a dataset of 87 authors with 500 characters message blocks. Equal error rate was 14.35%.

There are limited numbers of conducted researches related to Twitter. Natural language processing (NLP) in Twitter is dealt with by Lake (2010) for data extraction issue. This study did not investigate author identification; instead, related issues are directed like data structure. The study of Inches and Crestani (2011) also examined Twitter by data mining in text aspect. The work conducted by Dietrick et al. (2012) investigated gender identification in Twitter while the work of Bergsma et al. (2012) investigated automatic language identification.

Even though the study is limited to 3 authors, Sousa et al. (2011) in their study focused on author identification in Twitter. Green and Sheppard (2013) examined a series of 15 experiments involving up to 12 authors, with expanded feature sets and SVM was used. 92% success was reached with two authors but in other experiments results were around 40%. Mikros and Perifanos(2013) used author's multilevel n-gram profile for a dataset of 10 Greek twitter users with 12973 tweets and sizes of 25 to 100 words. With support vector machines they had success around 85%. Roy et al. (2013) tested a new concept of author's signature and a flexible pattern feature for tweets. They informed that their system obtains 6.1% improvement over the current state-of-the-art.

Jenny et al. (2014) used Facebook post, average 20.6 words as dataset for checking if user is authenticated or not among 30 users. SVM Light was used with 233 features and 12 tests were done. Success is between 87% and 98.6%.

Albadarneh et al. (2015) used big data collection of tweets from 20 authors in Arabic language. They implemented Naive Bayes classifier and accuracy was 61.6%.

5. SUPPORT VECTOR MACHINES

Support Vector Machines (SVMs) has taken attention to on current regarding the learning community (V.N. Vapnik, 1998). A SVM refers to a hyper plane as in its simplest linear form by distinguishing a number of positive examples from a number of negative examples with maximum interclass distance, the margin, in other words. Figure 1 represents such kind of hyper plane with the associated margin.

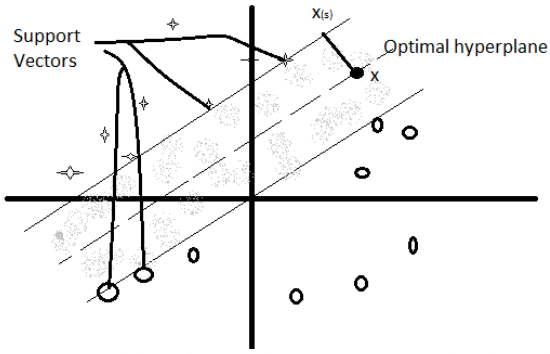

Figure 1. The idea of an optimal hyperplane for linearly separable patterns.

The formula used for the output of a linear SVM is as follows;

$$u = w * x + b \quad (1)$$

where w means the normal vector to the hyper plane, and x means the input vector. The margin is defined through the distance of the hyper plane to the nearest of the positive and negative examples.

It is important to know that the hyper plane is only determined by the training instances xi on the margin, meaning the support vectors. Surely, it cannot be said that all problems are linearly separable.

Regarding the computational learning theory, the structural risk minimization principle is the essential of support vector machines. Its aim is to detect a model for which the lowest true error can be assured. The model in which a bound on the true error is approximately reduced to the smallest degree with the control of the model complexity (VC-Dimension) is what an SVM determines. Over-fitting is departed from as the basic problem regarding semi-parametric models. Through mapping the input space into so high-dimensional feature space chosen a priori, nonlinear models can get extended to by the SVM. So, the optimal separating hyper plane is formed within this space (B.E. Boser, 1992).

The SVM is applied by Joachims to classify the text into various topic categories. Word stems were what he applied regarding the features. He claimed that each feature occurs at least three times in a text for constructing statistically significant features. A transductive SVM for text categorization was utilized by Joachims that makes the usage of the information in unlabeled training data possible.

6. RESEARCH METHODOLOGY

First step is data collection. Dataset was collected from Twitter for several authors. Four criteria were decided for author and message selection to narrow the scope and ensure optimal data availability. Criteria were:

- Users who frequently tweet
- Mainly new messages (instead of re-tweets or quotes)
- Individual authorship instead of group or corporate users

After choosing authors that meet these criteria, raw data was collected. Nvivo was used for collecting data. Raw data was preprocessed to remove messages that will not be useful in learning or testing process and to obtain the same structural data. Removing messages with less than two words or messages containing re-tweets are some form of preprocessing.

Second step is feature extracting. Four types of features were integrated into feature set and used for e-mail authentication. (Zheng et al, 2006) Adopted features are: Lexial features, syntatic features, structural features and content-specified features.

Since manual feature extraction is time consuming, we used an automated feature extraction program that was written in Java language. We had a large dataset with hundreds of features after the process. It is quite expensive to use high dimensional data set. It is important to select optimal features subsets. In this research we chose first 16 features (see Table 1) and these features were used to create input clusters. After choosing the features, dataset was normalized.

Table 1. Feature Set

# of character	# of word	# of sentence	# of comma
# of lower case letter	Average word length	# of upper case letter	# of short words
# of A letter	# of B letter	# of C letter	# of D letter
# of E letter	# of F letter	# of G letter	# of H letter

Third step is building the appropriate tool for classification. Author authentication model was designed by using support vector machine.

We used Wolfram Mathematica as software to create our tool.

Forth step is data processing. The dataset was split into two subsets. First subset is called training set and is used to train classification model. Second subset is called testing set and is used to validate the prediction ability of the generated model. Network is applied to train and build the prediction model before applying on the test set data. Training and testing process needs to be done iteratively to develop a successful authorship authentication model.

Last step is for evaluating of classification accuracy. After verifying the performance of classifier by the testing set, the model was used to authenticate the authorship of unknown online messages.

Standard accuracy measure (percentage of correctly classified text) was used in many researches. We also used this measurement to show our results.

7. A TEAM OF SUPPORT VECTOR MACHINES FOR AUTHORSHIP AUTHENTICATION

The support vector machine technique does not require much configurations compares to the neural network technique but choosing kernel is an important and mandatory task to do.

Support vector machines can refer different kind of activation functions such as sigmoid and RBFs. The activation function for the hidden layer of these machines is referred to as the inner product kernel, $K(x_i, x) = \varphi(x_i)$.

The support vectors are composed as the centers in radial base functions with the kernel equal to the activation function.

To transform linearly non separable data to a higher dimensional space where transformed data is linearly separable, Gaussian type radial base function is employed in our study:

$$\varphi(x) = \exp\left(-\frac{\|x-c_i\|^2}{2\sigma^2}\right) \quad (2)$$

where c_i is the vector represents the function center and σ is parameter affecting the spread of the radius.

The expansion of the kernel $K(x_i, x)$ permits us to construct a decision surface that is nonlinear in the input space.

Support vector machines have really good performance to classify two classes. They do not have same performance level for multiple classifications. Therefore we trained 10 support vector machines to classify 5 authors. The types to classify first author were {{1, 2}, {1, 3}, {1, 4}, {1, 5}} where each number represents one of the authors. For each author we trained SVMs as the first one.

To maximize objective function of the support vector machine, built-in function NMaximize of Wolfram Mathematica is employed. NMaximize function returns the values of weights α_i, i=1… 2m where m is the size of the training set from each of the two classes.

From each author 20 tweets were chosen randomly for training, testing and validating data respectively.

Support vector machines, that are trained to classify classes, are trained to distinguish between class i, and class j can also be used to distinguish between class j, and class i. We also added ten SVMs to distinguish between classes i, and class i which always vote zero. Therefore we create a classifier team that is a combination of 25 support vector

machines. We called this team as competing team that would vote for input to choose the correct author.

When a validation data enters into the classifier, a 5×5 decision matrix is created. The entered data is sent to the team and asked to classify. After classification of the entered data set done, the percentages are record to the row of the author. The matrix is filled for each row in that way.

8. RESULTS AND DISCUSSION

In this research, we proposed a team of support vector machines to identify authorship of Twitter messages. Firstly we used 20 messages from each of the 5 authors for evaluating effectiveness of the system. Classification accuracies are as in the confusion matrix below:

$$\begin{pmatrix} 70 & 15 & 15 & 0 & 0 \\ 15 & 35 & 30 & 15 & 5 \\ 30 & 5 & 30 & 25 & 10 \\ 5 & 15 & 10 & 60 & 10 \\ 5 & 5 & 10 & 5 & 75 \end{pmatrix}$$

Diagonal of the matrix shows us the success of the each author:

$$\{70,35,30,60,75\}$$

Average success is 54%. Fifth author's accuracy is achieved around 75%, and the rest is below that one.

Then we used 40 messages from each of the 5 authors for evaluating effectiveness of the system. Classification accuracies are as in the confusion matrix below:

$$\begin{pmatrix} 72 & 5 & 8 & 8 & 8 \\ 8 & 28 & 35 & 12 & 18 \\ 0 & 30 & 38 & 20 & 12 \\ 12 & 10 & 5 & 58 & 15 \\ 2 & 10 & 28 & 32 & 28 \end{pmatrix}$$

Diagonal of the matrix shows us the success of the each author:

$$\{72,28,38,58,28\}$$

Average success is 45%. First author's accuracy is achieved around 72%, and the rest is below that one.

This research was done for authorship identification of Twitter messages. Success of identifying validation data was calculated as total success of each tweet group of the authors (20/40 messages), not separately for each tweet. This is the reason for achieving low results.

Also results showed that increasing the number of tweets did not help to increase the success. Another finding is used feature set kept some authors' style and results are affected from that. In the future, we need to improve verification accuracy by choosing optimal feature sets.

REFERENCES

Abbasi, A. and Chen, H. (2005) "Applying Authorship Analysis to Extremist-Group Web Forum Messages," the IEEE Computer Society, pp. 1541-1672.

Alaa El-Halees, (2009) Filtering Spam E-Mail from Mixed Arabic and English Messages: A Comparison of Machine Learning Techniques, The International Arab Journal of Information Technology, vol. 6, no. 1.

Al-Badarneh, J. , Talafha, B. B. , Al-Ayyoub, M. and Benkhelifa, E. (2015) Using Big Data Analytics for Authorship Authentication of Arabic Tweets, 8th IEEE/ACM International Conference on Utility and Cloud Computing, pp. 448-452.

Altheneyan, A. S. and El BachirMenai, M. (2014) Naive Bayes classifiers for authorship attribution of Arabic texts, Journal of King Saud University – Computer and Information Sciences, vol. 26, pp. 473–484.

Anchal and Sharma, A. (2014) SMS Spam Detection Using Neural Network Classifier, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, no. 6.

Azarbonyad, H., Dehghani, M. and Marx, M. (2015) Time-Aware Authorship Attribution for Short Text Streams, in SIGIR'15.

Bogomolny, A. (2016) "Benford's Law and Zipf's Law." [Online]. Available: http://www.cut-theknot.org/do_you_know/zipfLaw.shtml

Bordes, A., Glorot, X., Weston, J. and Bengio, Y. (2012) Joint Learning of Words and Meaning Representations for Open-Text Semantic Parsing, in Proceedings of the 15th International Conference on Artificial Intelligence and Statistics(AISTATS) 2012, vol. 22.

Brocardo, M. L., Traore, I. Saad, S. and Woungang, I. (2013) Authorship Verification for Short Messages Using Stylometry, in Proc. of the IEEE Intl. Conference on Computer, Information and Telecommunication Systems (CITS).

Brocardo, M. L., Traore, I. Saad, S. and Woungang, I. (2014) Verifying Online User Identity using Stylometric Analysis for Short Messages, Journal of Networks, Vol. 9, No. 12.

Can, M. (2012) Teaching Neural Networks to Detect the Authors of Texts Using Lexical Descriptors, Southeast Europe Journal of Soft Computing, vol. 1, no. 1.

Chaski, C. E. (2005) Who's At the Keyboard? Authorship Attribution in Digital Evidence Investigations, International Journal of Digital Evidence, vol. 4, no. 1.

Chen, Z. Huang, L., Yang,W., Meng,P. and Miao,P. More than Word Frequencies: Authorship Attribution via Natural Frequency Zoned Word Distribution Analysis.

Clark, J., Koprinska, I. and J. Poon, J. (2003) A Neural Network Based Approach to Automated E-mail Classification, in Proceedings. IEEE/WIC International Conference, pp. 702 – 705.

Freeman, J. A. and Skapura, D. M. (1991) Neural Networks Algorithms, Applications, and Programming Techniques. Addison-Wesley Publishing Company.

Hadjidj, R., Debbabi, M. Lounis, H., Iqbal, F., Szporer, A.and Benredjem, D. (2009) Towards an Integrated E-mail Forensic Analysis Framework, Digital Investigation 5, pp.124-137.

Haykin, S. (1999) Neural Networks, A comprehensive foundation. 2nd Edition, New Jersey: Prentice Hall International.

Holmes, D. (1998) The evolution of stylometry in humanities scholarship, Literary and Linguistic Computing, vol. 13, no. 3, p. 7.

Holmes, D. I. and Forsyth, R. S. (1995) The Federalist Revisited: New Directions in Authorship Attribution, Oxford Journals, vol. 10, no. 2, pp. 111-127.

Holmes, D. I. (2016) Stylometry: Its Origins, Developments and Aspirations. [Online]. Available: http://www.cs.queensu.ca/achallc97/papers/s004.html

Howedi, F. and Mohd, M. (2014) Text Classification for Authorship Attribution Using Naive Bayes Classifier with Limited Training Data, Computer Engineering and Intelligent Systems, vol.5, no.4.

Koppel, M. and Winter, Y. (2014) Determining if two documents are written by the same author, Journal of the Association for Information Science and Technology, vol. 65, pp. 178–187.

Koppel, M., Schler, J., Argamon, S. and Messeri, E. (2006) Authorship Attribution with Thousands of Candidate Authors," SIGIR'06.

Kourtis I. and Stamatatos, E. (2011) Author Identification Using Semi-supervised Learning, CLEF 2011 Labs and Workshop.

Layton, R., Watters, P. and Dazeley, R. (2010) Authorship Attribution for Twitter in 140 Characters or Less, Second Cybercrime and Trustworthy Computing Workshop.

Li, J. S., Monaco, J. V., Chen, L.C. and Tappert, C. C. (2014) Authorship Authentication Using Short Messages from Social Networking Sites, in Proceedings of the 2014 IEEE 11th International Conference on e-Business Engineering, pp. 314-319.

Luyckx, K. and Daelemans, W. (2011) The effect of author set size and data size in authorship attribution," Literary and Linguistic Computing , vol. 26, no. 1.

Mikros, G. K. and Perifanos, K. A. (2013) Authorship Attribution in Greek Tweets Using Author's Multilevel N-Gram Profiles, in Analyzing Microtext: Papers from the 2013 AAAI- Association for the Advancement of Artificial Intelligence Spring Symposium.

Narayanan, A., Paskov, H. , Gong, N. Z. , Bethencourt, J. et al. (2012) On the Feasibility of Internet-Scale Author Identification, 2012 IEEE Symposium on Security and Privacy, pp. 300 – 314.

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

Sanderson, C. and Guenter, S (2006) Short Text Authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking: An Investigation, in Proceedings of International Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 482–491.

Silva, R. S., Laboreiro, G., Sarmento, L.,Grant, T., Oliveira, E. and Maia,B. (2011) 'twazn me!!! ;('Automatic Authorship Analysis of Micro-Blogging Messages, Natural Language Processing and Information Systems, vol. 6716, pp. 161-168.

Stamatatos, E. (2009) A Survey of Modern Authorship Attribution Methods, Journal of the American Society for Information Science and Technology , vol. 60, no. 3, pp. 1-28.

Vorobeva, A.A. (2016) Forensic Linguistics: Automatic Web Author Identification, Scientific and Technical Journal of Information Technologies, Mechanics and Optics, vol. 16, no. 2, pp. 295–302.

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.