Stylometric Analysis for Authorship Attribution on Twitter

Mudit Bhargava, Pulkit Mehndiratta, and Krishna Asawa

Jaypee Institute of Information Technology
muditbhargava09@gmail.com,
{pulkit.mehndiratta,krishna.asawa}@jiit.ac.in

Abstract. Authorship Attribution (AA), the science of inferring an author for a given piece of text based on its characteristics is a problem with a long history. In this paper, we study the problem of authorship attribution for forensic purposes and present machine learning techniques and stylometric features of the authors that enable authorship to be determined at rates significantly better than chance for texts of 140 characters or less. This analysis targets the micro-blogging site Twitter[1], where people share their interests and thoughts in form of short messages called *"tweets"*. Millions of *"tweets"* are posted daily via this service and the possibility of sharing sensitive and illegitimate text cannot be ruled out. The technique discussed in this paper is a two stage process, where in the first stage, stylometric information is extracted from the collected dataset and in the second stage different classification algorithms are trained to predict authors of unseen text. The effort is towards maximizing the accuracy of predictions with optimum amount of data and users under consideration.

Keywords: Online Social Media, Twitter, Authorship Attribution, Machine Learning Classifier, Stylometry Analysis.

1 Introduction

In recent years, authorship attribution of anonymous messages has received notable attention in the cyber forensic and data mining communities. During the last two decades this technique has extended to computer facilitated communication or online documents (such as e-mails, SMS, Tweets, instant chat messages etc) for prosecuting terrorists, pedophiles, and scammers in the court of law. In the early 19^{th} century it was considered difficult to determine the authorship of a document of fewer than 1000 words. The number decreased significantly and by the early 21^{st} century it was considered possible to determine the authorship of a document in 250 words. The need for this ever decreasing limit is exemplified by the trend towards many shorter communications techniques like Twitter, Facebook[2], Short Message Services (SMS) etc.

[1] https://twitter.com/

[2] https://www.facebook.com/

V. Bhatnagar and S. Srinivasa (Eds.): BDA 2013, LNCS 8302, pp. 37–47, 2013.

Authorship Attribution of online documents is different from the authorship attribution of traditional work in two ways. Firstly, the online documents or text collection is mostly unstructured, informal and not necessarily grammatically correct as compared to literature, poems and phrases which are syntactically correct, very well structured and elaborative in nature. Secondly, for a single online document the number of authorship disputes are far more as compared to traditional published documents, that is because one of the challenges with authorship attribution in this case is scarcity of standardized data to test the accuracy of results.

Online social networks (OSN) like Twitter, Facebook, Linkedin[3] give a new dimension to the authorship attribution all together. These online networks provide effective and fast means of communication for the conduct of any criminal activity by anonymous users. User may use screen names or pen-names on these sites, while others may not provide the correct identification information with the accounts. On top of it, a single user can create multiple profiles on these online social networks. Anonymity poses bigger threats for law enforcement agencies in tracking the identity of these users. This paper is an effort towards addressing the problem of authorship attribution for an online social network Twitter. Twitter has surged popularity in recent years and now reports that it has over 500 million user base which share almost same number of messages (called *tweets*) per day [4].

This paper focuses on the problem of identification of the original author for a given tweet from a list of suspected authors for it using stylometric information. Various stylometric features have been taken into consideration for the training and later testing purposes of the machine learning algorithms such as Support Vector Machine (SVM) classifier. The experiments have been conducted on different datasets using the mentioned approach and in the last ideas for the future work have been discussed

2 Background

There has been an unprecedented growth in the amount of short messages that are shared worldwide everyday. Status messages on Twitter and Facebook, comments on the YouTube[4] and news blogs show a clear trend on using short messages for daily communication on the internet. With the advent of free text messaging mobile applications like Whatsapp[5], Viber[6] and the existing SMS, millions of short messages are shared through mobile phones.

A similar trend has also been seen in cyber-crime, where fraudulent activities like identity frauds and cyber-bullying are usually done with shorter messages such as fraud e-mails, forum posts, on Facebook and Twitter, as well as many other websites. Thus, it is of utmost importance to come up with some technique

[3] `https://www.linkedin.com/`

[4] `https://www.youtube.com/`

[5] `http://www.whatsapp.com/`

[6] `http://www.viber.com/`

or method to identify the authors of the various short messages posted daily on these websites. However, there are several stylometric features comprising of lexical, syntactic, structural, content-specific and idiosyncratic characteristics. Many studies have been proposed which consider stylometric features for performing the authorship attribution but still a lot of work has to be done for messages of length as low as 140 characters or less.

2.1 Related Work

In field of Stylometry, linguistic characteristics of a language are studied to gain knowledge about the author of the text. In [1], Abbasi *et al.* have talked about the issue that how the anonymity hinders the social accountability and tried to identify the author based on his/her writing style. Similarly in [2], authors have tried to apply mining techniques to identify the author of any particular e-mail. Koppel *et al.* [3] have tried to automatically categorize the written text and find out the gender of the author. Word frequencies, word length, number of sentences are considered important stylometric features as mentioned in [5].

The decrease in the size of the document has broadened the area of applications available. Early work on authorship attribution was focused on large documents like Federalist papers [8], but with recent developments in the field, attribution of authorship has even become applicable to blog posts [7] and short forum postings [6, 12].

R. Rousa Silva et.al [19] have worked on authorship attribution using stylistic markers for tweets written in Portuguese. Their analysis shows how emoticons and short messages specific features dominate over traditional stylometric features to determine the authorship of tweets. The final results show significant success (i.e. F Score = 0.63) for 100 examples available from each author under consideration. However the numbers of authors under suspect have been limited to 3 at a time.

Also, a recent technique involved using Probabilistic Context-Free Grammars (PCFG) for the attribution of author of a document [9]. Both lexical and syntactic characteristics were taken into consideration to capture an author's style of writing. The corpus that was included had data from different fields such as poetry, football, business, travel, and cricket. The method involved, first implementing a PCFG for all authors separately and then build a training model using the grammar for performing the classification. This model was combined with other models (bag-of-words Maximum Entropy classifier and n-gram language models) which captured lexical features too. For the dataset related to cricket, 95% of the total instances were correctly classified.

The next section provides an overview of the data set, how the tweets are collected, cleaned, pre-processed and clustered/grouped under one author. Section 4 discusses about the methodology used for the authorship attribution. In section 5 we discuss about the various results that we achieve including the accuracy of the experiments conducted. Finally, section 6 provides the conclusion of the experimental design used for the attribution of authors.

3 Data Set

Data always plays a critical role in authorship attribution. For performing the same on Twitter, an author attributed tweet dataset is required. Standard twitter corpora contain multiple tweets from multiple authors. Moreover, twitter terms of use do not allow distribution of tweets. Twitter corpora are collection of user IDs and tweet IDs. Downloading the content using automated scripts that accompany these corpora is a time consuming task, as twitter servers cannot be hit hard with too many requests at once. Owing to these constraints, we create our own dataset of author classified tweets using a Twitter client application that randomly collects public statuses using Twitter streaming application programming interface (API). A python based twitter corpus tool from [14] returns a random sample of about 5000 public statuses and stores them to disk in Java Script Object Notation (JSON) format.

Requirement is for the users/authors who would undergo the stylometric analysis. Choosing these users for experiments is a three step task. From the 5000 public statuses that have been collected, a list of unique authors is generated. Next, the requisite number of users are selected from the list randomly. Lastly, 300 most recent public statuses of these selected authors are streamed using GET statuses/user_timeline API from twitter [15].

We require users to have a certain threshold number of tweets (discussed in Section 4) and their language of profile and tweets to be English. Hence, if a selected user doesn't meet this criteria, we randomly select another user from the list of unique authors. Tweets streamed are parsed for their text or 'tweet' content and twitter specific features like 'hashtags', 'usermentions' and 'embedded urls'.The use and impact of these features is discussed in the further sections.

4 Methodology

Stylometric analysis on tweets is similar to those done on other forms of short texts such as web-forum posts or online instant messaging chats. They are informal and similar in structure and syntax [13]. An exhaustive feature set considering stylometric information is built for our experiment, however with an assumption that the authors unconsciously follow a specific pattern and are consistent in their choices [12]. Various broad categories of the features are as follows:

1. Lexical Features:
   (a) Total number of words per tweet
   (b) Total number of sentences
   (c) Total number of words per sentences
   (d) Frequency of dictionary words
   (e) Frequency of word extensions
   (f) Lexical diversity
   (g) Mimicry by length
   (h) Mimicry by word

Even though tweets are really short texts, users would manage to write them using dictionary words and framing proper sentences (features b, c, d). Feature (e) looks out for authors who would have a habit of extending the words by repetition of the last or intermediate letter, for example ’hiiii!!!’, ’heyyy!’, ’meee’ etc. We learn from instant messaging [13] that how multiple messages from the same user are related in terms of vocabulary, length, attitude and that the user mimics his own style in each messages he sends. Though instant messages are different from tweets in numerous ways, we see quantifying mimicry as the ratio of length [13] in two chronological tweets (feature (g)) and as the number of common words in those two pieces of text, provide valuable information about the writing styles of the author. For example in the following tweets, ’Actually, I have to go out with friends today’, ’Actually, I was a little busy today’, ’Oh! Cool. That rocks!’, ’Oh cool, the plan is on’ we can see that how the author has a habit of using the words such as ’actually’ or ’cool’ quite often.

2. Syntactical Features:
   (a) Total number of beginning of sentences (BOS) characters capitalized
   (b) Number of punctuation per sentence
   (c) Frequency of words with all capital letters normalized over number of words
   (d) Frequency of alphanumeric words normalized over number of words
   (e) Number of special characters, digits, exclamation and question marks
   (f) No of upper case letters
   (g) Binary feature indicating use of quotations

   There would be users ho would write their entire messages in capitalized letters, HELLO ARE YOU THRE? (feature (c)) and many users would condense words with the combination of characters and digits (like tonight becomes 2night or tomorrow becomes 2morrow, feature (d)) to fit in their entire content in the specified character limit. Features from (e)-(f) cover messages types where a different amount of special characters are used like ”∗, #, %, ˆ”etc., words are irregularly capitalized (eg. heLLo) or have too many question marks or exclamations (e.g. What??? or hi!!!). Feature (g) takes in account styles of those users who like to post popular quotes or quote other users.

3. Features specific to tweets:
   (a) Binary feature indicating if the tweet is a re-tweet
   (b) Number of hash-tags normalized over number of words
   (c) Number of user mentions normalized over number of words
   (d) Number of URLs normalized over number of words

There are features that are unique to twitter posts only. Re-tweet is sharing of a tweet originally composed by another author, hash-tags are used to convey the subject of a tweet (#sports,#now playing etc.), user mentions tag other users/sends them replies (@user1, @user2) and URLs are attached to share pictures, videos etc. Users would often bloat their tweets with hash-tags or user-mentions or have a very high frequency of re-tweeting. These features try and cover such stylometric information.

4. Other helpful features:
   (a) Frequency of Emojis[7]
   (b) Number of Emojis per word
   (c) Number of Emojis per character

Most emoticons (:), :P, :/) in data are converted to emojis and have a special unicode representation, making it difficult to detect them using syntactical features. Hence they need to be accounted for differently.

4.1 Grouping of Tweets

Even though a tweet can be at most 140 characters long, many authors use even lesser characters to express themselves. For example tweets like 'Good Morning followers' or replies like , '@someone Thanks!' are just 2-3 word long or 16-22 characters long. Such tweets would not have much stylometric information to contribute. One solution therefore is to remove such tweets from our dataset and work on texts with greater number of words. However in doing so, we loose out on important information about the style and traits of the author. Also if a user has a tendency of sharing only very short messages, with just 5-6 words, our system would fail to make predictions about such an author. Hence to overcome this challenge we group various tweets and increase the text size under consideration. Now to analyze patterns over a group of tweets rather than one single tweet is easier and fruitful. So we are mapping a group of tweets to its author rather than a single tweet and this is because some tweets might be excessively small as described above. The assumption here is that the tweets grouped together are from the same account and only one user maintains the twitter account under consideration. We also did a quantitative analysis, by bunching tweets in different group sizes. An overview of results obtained are summarized and illustrated in Table. 1 and Fig.1. Accuracy (%) is number of authors correctly classified. Group size greater than 4 provide acceptable results, but noticeably a group of 8-10 performs the best. A detailed analysis has been done in the further sections.

4.2 Experiments

As discussed in previous sections, a prerequisite to train machine learning algorithms and thereby performing text classification on the data collection is

[7] Japanese term for pictures characters and emoticons.

Table 1. Variation of accuracy with different tweet grouping sizes

Groupsize	Accuracy%
1	7.81
2	70.09
4	81.23
6	80.25
8	84.73
10	91.11

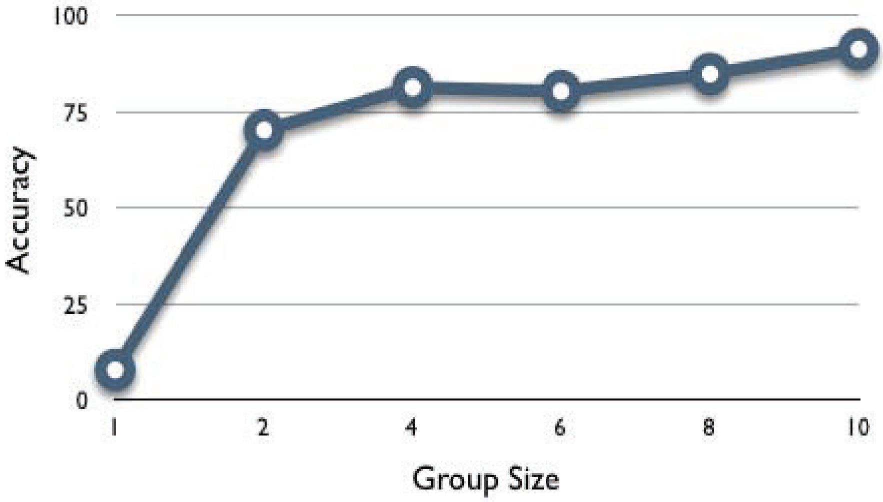

Fig. 1. Variation of accuracy with different tweet grouping sizes

selection of valid stylometric features. However these features need to be quantized for the collected data set. To perform that, a script has been generated that extracts features from the input tweets and groups them according to section (4.1) for further analysis. The script uses regular expressions and standard functions from the Natural Language Tool Kit (NLTK) [16] for python. The features and corresponding author labels will now be used to perform supervised machine learning using Support Vector Machine (SVM) [17].

4.3 Classification

Classification includes scaling of data, choosing the right SVM kernel, calculating best kernel parameter values and finally, testing the designed model for results. To avoid attributes with large numeric ranges to dominate over the ones with smaller ranges, it is essential to perform basic scaling of the attributes. So, if a feature has a value from [-10 to 10] it gets scaled to [-1 to 1]. Also, from various SVM kernels available it is important to choose the right one for best results. For this experiment, we pick the Radial Basis Function (RBF) kernel, which is a better choice over the linear, sigmoid or the polynomial kernels for several reasons in this scenario. RBF can handle cases where the relationship between class labels and features is non linear, this is because it non-linearly maps samples into a higher dimensional space. RBF also has fewer hyper-parameters affecting complexity of model selection and fewer numerical difficulties in comparison to polynomial and sigmoid kernels [10]. Since, features taken under consideration for this experiment range between [23-25] and are not very large in comparison to the number of instances (varying from 200 to a 1000), the RBF is the best kernel choice. There are two parameters for the RBF kernel: C and γ. The libsvm package [18], used for analysis, performs the grid-search on C and γ using cross validation. It is a naive yet efficient approach that picks the values for the parameters on the basis of best cross-validation accuracies. These parameters are fed into the libsvm script, which further classifies each input in the testing set by using the one versus all technique.

As illustrated in Table. 2 , we vary both the number of authors and their number of tweets to check how our classification model performs in all possible situations. The results discussed are for a group size of 10 tweets, where grouping is done as described in Section (4.1). The accuracy is the number of authors correctly classified by the SVM classifier. The quality of classification is defined by the F-Score which is calculated by the standard formula -

$$\frac{(2*Precision*Recall)}{(Precision+Recall)}$$

where precision and recall are as:

$$\text{Precision}= \frac{(True\ positives)}{(True\ positives+False\ positives)}$$

$$\text{Recall}= \frac{(True\ positives)}{(True\ positives+False\ negatives)}$$

The number of tweets varies from 200 - 300 and the number of authors varies from 10 - 20. The number of authors and tweets has been kept low with the following assumptions and constraints.

If we have fewer than 200 tweets, and we group them as explained earlier, we will end up with very few training instances per author. In a maximum grouping size of 10, we have 20 instances per author, which despite being less performs decently as it will be discussed in next section. If we have more than 300 tweets, we are asking for too much of data for one user, which might not always be available. Moreover, 300 most recent tweets give us sufficient stylometric information and we see in the next section that how the increase this number affects the performance of classification.

The number of authors have been kept low because, we consider that the list of suspected authors, who are under the scanner have been nailed down upon by other means like conduct of criminal/ non-criminal investigations. This is also the basis for the work done in [11] and [19].

Table 2. Results considering all features in a tweet grouping size of 10

Tweets	Users	Accuracy%	Precision%	Recall%	F-Score%
200	10	81.42	90.22	81.42	85.59
200	15	83.80	89.73	91.42	90.56
200	20	56.42	66.20	62.85	64.48
250	10	77.77	84.14	77.70	80.79
250	15	91.11	95.16	94.44	94.79
250	20	59.40	71.88	71.11	71.49
300	10	75.45	78.73	75.45	77.05
300	15	84.84	84.84	93.18	88.64
300	20	64.54	64.54	77.93	77.13

5 Results

Given just 200 tweets per author, and 10-15 suspects, we obtain an F-score in the range of (85.59% to 90.56%). However, if the number of suspects is increased to 20, the F-score drops drastically to a low value of 64.48%. Evidently in case of lesser number of tweets, we need to narrow down on our list of suspected authors. For 250 tweets per author, again best results have been achieved with 15 authors, where the F-score reaches 94.79%. With increase in data, we have an increase in F-Score for 20 authors (from 64.48% to 71.49%) indicating how an increase in content might be required with more number of users. With 300 tweets, again our best F-score is 88.64% with 15 authors. The F-score for 20 authors further increases with increase in number of tweets under consideration.

Table 3. Analysis of accuracy varying with features under consideration

Features	Accuracy%
All	91.11
Syntactica + Lexical + Tweet Specific	85.09
Syntactical + Lexical + Others	83.38
Syntactical + Tweet Specific + Others	83.38
Syntatical	72.20
Lexical + Tweet Specific + Others	62.74

There is notable decrease in the best obtainable accuracy as we increase the number of tweets. Since our tweets are collected in chronological order one possible inference from this observation is that the style of the author varies over time. Just similar to the case, when style of writing formal e-mails would differ from the ones to our friends and family, the style of writing tweets may differ over time because of the different topics that people are talking about over the time on Twitter. For example, if it's the Premier League season or there is an on-going cricket series, we would see a bias towards sports content amongst sport enthusiasts. This being a data dependent application, results are bound to vary over different data sets.

Table. 3 shows how different category of features make an impact on the accuracy of classification. Results are compiled for 250 tweets per author and 15 suspected authors. Non consideration of twitter specific features reduce the accuracy by 7.7%, however there is only a 6% decrease when features related to emoticons are eliminated from the analysis. Removal of syntactical features have a very strong impact on classification accuracy, reducing it by 28%. Though necessary, they are not sufficient for our experiment. Considering only syntactical features results in a low accuracy rate of 72.2 %.

As discussed in section 2, [19] also uses stylometric features for tweet authorship attribution. The former study requires 100 tweets per author and considers only 3 suspected authors at a time to achieve at most an F-Score of 0.63; by adding just a 100 more tweets our analysis can be extended to 10 suspected authors at a time and provide significantly better prediction accuracies (F-Score = 0.85, for 200 tweets, 10 users). With grouping of tweets in sets of 2-10 tweets per author and using the one versus all classification technique with SVMs, the techniques discussed in this paper and [19] are two different ways of using stylometric features for twitter authorship attribution; each having their own important contributions to the domain.

6 Conclusions and Future Work

Our study of authorship attribution for twitter shows interesting initial results. We have achieved a precision of up to 95.16 % , and a F-Score of up to 94.79% over a data set that is collected with no bias towards any specific content, user or geographical area. We also see how grouping of tweets together has an impact on author based tweet classification. It can be concluded that 200-300 tweets per author and list of 10-20 such suspected authors form a practical data set for analysis. The listed features, the SVM classifier with the RBF kernel and its optimum parameters values, form a good model for stylometric analysis of tweets from twitter.

In future, we would like to reduce the number of tweets per author required for defining a stylometric pattern. Also, a few points in the obtained results require detailed reasoning. These may become clearer by using an elaborate data set and performing further experiments. Other important tasks that we plan to undertake include, increasing the number of suspected authors under consideration and adding more precise features that would uniquely identify the tweets in question.

References

1. Abbasi, A., Chen, H.: Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Transactions on Information Systems (March 2008)

2. de Vel, O.: Mining e-mail authorship. In: ACM International Conference on Knowledge Discovery and Data Mining (KDD) (2000)
3. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), 401–412 (2002)
4. Twitter report twitter hits half a billion tweets a day (October 26, 2012), `http://news.cnet.com/8301-1023_3-57541566-93/report-twitter-hits-half-a-billion-tweets-a-day/`
5. Holmes, D.I.: The evolution of stylometry in humanities scholarship. Literary and Linguistic Computing 13(3), 111–117 (1998)
6. Abbasi, A., Chen, H.: Applying authorship analysis to extremist-group web forum messages. IEEE Intelligent Systems 20(5), 67–75 (2005)
7. Mohtasseb, H., Lincoln, U., Ahmed, A.: Mining Online Diaries for Blogger Identification. In: Proceedings of the World Congress on Engineering (2009)
8. Mosteller, F., Wallace, D.L.: Inference in an authorship problem. Journal of the American Statistical Association 58(302), 275–309 (1963)
9. Raghavan, S.: Authorship Attribution Using Probabilistic Context-Free Grammars. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL (2010)
10. Hsu, C.-W., Chang, C.-C., Lin, C.-J.: A Practical Guide to Support Vector Classification. Department of Computer Science, National Taiwan University, Taipei, Taiwan (2010)
11. Malcolm Walter Corney, Analysing E-mail Text Authorship for Forensic Purposes. Queensland University of Technology, Australia (2003)
12. Pillay, S.R., Solorio, T.: Authorship Attribution of web forum posts. APWG eCrime Researchers Summit (2010)
13. Cristani, M., Bazzani, L., Vinciarelli, A., Murin, V.: Conversationally-inspired Stylometric Features for Authorship Attribution in Instant Messaging. ACM Multimedia (October 29, 2012)
14. Twitter Corpus (2012), `https://github.com/bwbaugh/twitter-corpus`
15. Twitter (2013), `https://dev.twitter.com/docs/api/1/get/statuses/user_timeline`
16. Natural language Toolkit (2013), `http://nltk.org/`
17. Support Vector Machine (2000), `http://www.support-vector.net/`
18. Libsvm (2013), `http://www.csie.ntu.edu.tw/cjlin/libsvm/`
19. Sousa Silva, R., Laboreiro, G., Sarmento, L., Grant, T., Oliveira, E., Maia, B.: 'twazn me!!! ;(' Automatic Authorship Analysis of Micro-Blogging Messages. In: Muñoz, R., Montoyo, A., Métais, E. (eds.) NLDB 2011. LNCS, vol. 6716, pp. 161–168. Springer, Heidelberg (2011)