Authorship Recognition of Tweets: A Comparison Between Social Behavior and Linguistic Profiles

Madeena Sultana  
Dept. of Computer Science  
University of Calgary  
2500 University Dr NW, Calgary, AB, Canada  
Email: msdeena@ucalgary.ca

Padma Polash Paul  
University of Oxford  
Oxford, 0X1 3PA, U.K.  
Email: pppaul@ucalgary.ca

Marina Gavrilova  
Dept. of Computer Science  
University of Calgary  
2500 University Dr NW, Calgary, AB, Canada  
Email: mgavrilova@ucalgary.ca

Abstract—Authorship recognition from micro-blogs such as Twitter is a challenging task due to limitation of text length to 140 characters. However, identification of micro-blog authors is crucial in many cyber-crime investigations as well as in forensic applications. So far, traditional linguistic profiles such as Bag-Of-Words (BOW) and style-based markers have been investigated for identification of micro-blog authorship. The social interactive data in micro-blogs remained understudied for this purpose. In this paper, we examined authorship recognition based on the social interactions of users in Twitter and present a comparative analysis with BOW and style-based features. We obtained 97% recognition rate on a database of 70 Twitter users, which validates the superiority of using social interactive data compared to traditional linguistic profiles.

Index Terms—Social behavioral biometrics, authorship recognition, style-based features, bag-of-words, linguistic profiles, social behavior, social data mining, online social network.

I. INTRODUCTION

Automated authorship recognition is a subfield of Natural Language Processing (NLP) that identifies the author of an anonymous piece of writing by measuring linguistic similarity between the anonymous writing and a set of writing samples of the known authors [1], [2]. The basic assumption behind linguistic "writeprint" is that every author has a personal style of writing in terms of using vocabulary, word frequencies, order, style, grammar, punctuation, etc., which can distinguish one author from another. Linguistic writeprints have been successfully utilized for author identification of books [3], literature [4], scientific papers [5], news articles [6], and so on. Over the last decade, the new forms of written communication such as e-mail, blogs, web-forums, instant messages, Short Message Service (SMS), etc. have gained popularity due to the availability and cost reduction of Internet and communication devices. Authorship recognition has also been applied to e-mails [7], web-forums [8], blogs [9], chat conversations [10], SMS texts [11] in order to facilitate law enforcement, forensic applications, and cybercrime investigations.

Nowadays, the unprecedented rise of the Online Social Networks (OSN) provides users a massive platform to stay socially connected beyond geographic boundaries. Some OSNs provide social communication via short form of texts known as micro-blogs. Similar to other social media accounts (e.g., e-mail, web-forum accounts, etc.), the micro-blogging accounts of users are vulnerable to various security threats such as forgery, deception, vandalism, identity theft, fake accounts, cyberbullying, and so on. Therefore, authorship recognition of micro-blogs is often needed to investigate cybercrimes and prevent security threats. It is also possible to explore personality, gender, age, emotional state, etc. of persons by applying authorship recognition techniques on micro-blogs and social media contents [12], [13]. However, there are some unique challenges of micro-blog authorship recognition since micro-blogs

1) consist of terse text. The popular micro-blogging service Twitter has a limitation of length of only 140 characters.
2) are informal way of communication, which lacks syntactical forms of formal writing.
3) content is inconsistent and topics may vary over time.
4) contain highly unstructured data - a combination of web-links, hashtags, emoticons, special symbols, images are present along with texts.

Due to the aforementioned factors, identification of authors of micro-blogs using traditional linguistic profiles is not as effective as authorship attribution of larger and formal texts. Micro-blogs are intended for social communication with a larger community, where social interactions take place in replies, mentions, retweets, web-links, hashtags, etc. However, use of such social interactive data remained understudied for authorship recognition. In this paper, we propose to use social behavioral profile to determine authorship of micro-blogs. Recently, we conceptualized that individuals have idiosyncratic way of social interactions, which we denoted as ‘Social Behavioral Biometrics’ [14]. Our latest research revealed that social behavioral biometrics of users’ via OSN possess uniqueness and permanence over a certain period of time [15]. In this paper, we propose to create social behavioral profile of users using a subset of the social behavioral biometric features for micro-blog authorship recognition.

The novel contribution of this research is two-fold:

• We propose a novel way of mapping ambiguous micro-blog authors identity to previously known authors based on their Social Behavioral (SB) profile.
• We present a comparative analysis of micro-blog authorship recognition using SB profiles and two popular lin-
guistics profiles: Bag-of-Words (BOW) and style markers. The stability of these methods over time are also analyzed using 4 different sessions of tweet samples.

II. RELEVANT RESEARCH

A study on an Internet-scale dataset reported that authorship recognition methods based on linguistic stylometry tend to drop significantly with the reduction of text length smaller than 500 words [16]. However, the latest research showed that it is possible to recognize authors of micro-blogs by choosing an efficient feature set [1]. Some of the existing works on micro-blog authorship recognition are discussed below.

In 2010, Shalhoub et al. [17] proposed a method for automated author recognition from Tweets based on stylistic features. However, due to the limitation of 140 characters of tweets, the authors proposed to create the training samples from other pieces of writing of the authors such as books, blogs, etc. and match tweets with those. The use case considered in [17] was quite limited in the scope since it can only be applied if a user already has some other form of writing samples. During the same year, Layton et al. [18] proposed an authorship recognition method for Twitter based on Source Code Authorship Profile (SCAP) and n-gram. Unlike [17], Layton et al. [18] created both training and testing profiles from tweets and obtained 70% recognition accuracy on randomly selected 50 Twitter users. In 2011, Silva et al. [19] proposed to create stylometric model using 'emoticons', interjections, punctuation, abbreviations, etc. and used SVM classifier to determine the authorship of tweets. The highest recognition performance obtained by Silva et al. [19] was F-score=0.63 on 40 groups of 3 authors. Another study by MacLeod and Grant [20] showed that training with aggregated tweets produces better results than single tweet. In [20], grammar, lexis, punctuation-based stylistic features on candidate sets of 2, 5, 10, and 20 authors were used. A comparative analysis of frequency and style markers has been accomplished by Green and Sheppard [1] in 2013, which reported better performance of context free style markers than the vocabulary based Bag-Of-Words (BOW) method. Although 92.3% recognition rate on a candidate set of two authors was obtained in [1], the recognition rate dropped to 40% on a candidate set of 12 authors. A linguistic style-based author recognition method has been applied by Keretna et al. [21], where promising results were obtained on 30 Twitter users.

It can be noticed from the above discussion that the authorship recognition of tweets or micro-blogs has been investigated largely based on linguistic features, but the stability of the linguistic profiles over time have not been investigated. Most importantly, the essential part of tweets - interactive data such as replied, mentioned, retweeted friends, shared hashtags, web-links, etc., remained unexplored as a solution of authorship recognition problem. However, one of our very recent works showed that the authorship of the crowd-sourced contents (e.g., Wikipedia) can be predicted from the editing behavior of authors [22], which provides a good alternative of linguistic-based features. In this paper, we introduced authorship recognition method of micro-blogs using social interaction-based features called as social behavioral profiles. We also investigated the stability of the proposed method and linguistic profiles using 4 sessions of data collected over a period of four months.

III. METHODOLOGY

In Twitter, users communicate by posting real time micro-blogs called Tweets that provide information about user’ social behavior [14], [23]. Tweets comprises of texts as well as interactions to other users: replies, retweets, and mentions. Tweets directed to a specific user is either a reply or mention. Sharing other user’s tweet is called retweeting [24]. We also considered shared hashtags and web-links as social interactions since they represent user’s personal interests and affiliations shared with acquaintances. Hashtags are shorthand conventions of posts on the same topic [25]. In this paper, we created social behavioral profiles of Twitter users by combining these four interactive social features. A corpus of social behavioral profiles of Twitter users has been created, which can be learned by the authorship recognition system and used as reference to determine the identity of anonymous micro-blog authors. A flow diagram of the training and recognition phase of the proposed method is presented in Fig. 1.

The followings are some advantages of Social Behavioral (SB) profile over linguistic profiles:

- Social interaction-based features of SB profile are content independent. Therefore, changes in linguistic content of tweets do not affect the recognition performance.
- The SB profile features are not limited to 140 character tweets. A set of 50 tweets can contain reasonable amount of interactive features.
- SB profile features have well-structured forms (preceded by @, RT, #, http). Therefore, easy to parse from unstructured text.
- SB profile features do not depend on formal or informal forms of writing and applicable to any language.

The detailed description of the SB profile creation process is presented in the following subsection.

A. Proposed Social Behavioral Profile

At first, we extracted a list of replied, retweeted, and mentioned acquaintances from a set of 200 tweets of 70 users. We also extracted shared hashtags and domain names from tweets. Then, the following weighted feature vectors are created

Replied Friends: For each user in the corpus, we created a feature vector containing a list of unique acquaintances, whom the user replied in the last 200 tweets. Each unique name then followed by the log frequency of the occurrence of the item in the corpus. For instance, if there are $N$ replied friends in the last 200 tweets of a user, the ‘reply’ feature vector of the user contains $N < \text{item}, \text{weight} >$, where the $\text{item}$ is the replied friend and $\text{weight}$ is the log frequency weight of the $\text{item}$. The log frequency weight is calculated using Eq. 1 [26]:

$$W_t = 1 + \log(TF_{t,d})$$

(1)
where $TF_i$ is the Term Frequency of item $t$ in profile $d$.

Retweeted/Mentioned Friends: Similar to replied friend feature vector, we created retweeted friend feature vector by accumulating the list of all retweeted and mentioned acquaintances from the last 200 tweets of each user in the corpus. For total $N$ retweeted and mentioned friends in the last 200 tweets of a user, the 'retweet' feature vector of the user contains $N < item, weight >$, where the item is the retweeted/mentioned friend and weight is the log frequency weight of the item. The log frequency weight is calculated using Eq. 1.

Shared Hashtags: The shared hashtags in tweets often represent the user’s interest on some topics or events. Users share common hashtags or create their own. Similar to 'reply' and 'retweet' feature, we extracted all shared hashtags from the last 200 tweets of each user in the corpus. Since users share popular hashtags along with self-created hashtags in tweets, we assigned more weights on uncommon or user’s own hashtags, while reducing the weights of common hashtags. TF-IDF weights have been applied on hashtags. For total $N$ shared hashtags in the last 200 tweets of a user, the 'hashtag' feature vector of the user contains $N < item, weight >$, where the item is the hashtag and weight is the TF-IDF weight of the item in corpus. The TF-IDF weight is calculated using Eq. 2 [26]:

$$W_t = (1 + \log(TF_{i,d})) \times \log\left(\frac{N}{DF_t}\right)$$  \hspace{1cm} (2)

where $TF_i$ is the Term Frequency of each item $t$ in profile $d$, $N$ is the number of users profile in dataset, and $DF_t$ is the Document Frequency i.e. number of occurrences of the item $t$ in $N$ users’ profiles.

Shared Web-links/Domains: Lastly, we extracted a list of web-links shared by each user in the corpus. Since users can refer to different pages of same website, we considered only the domain names that the users are interested in. Similar to hashtags, some domains such as Youtube, Facebook, Instagram, Google, etc. are very common and shared by many users in tweets. Therefore, we applied TF-IDF weights (Eq. 2) to each domain names, which represent the degree of preference of each user to a specific domain. For total $N$ shared domains in the last 200 tweets of a user, the 'domain' feature vector of the user contains $N < item, weight >$, where the item is the domain name and weight is the TF-IDF weight of the item in corpus.

The proposed SB profile contains the aforementioned four interactive features: reply, retweet, hashtag, and domain. As shown in Fig. 1, during authorship recognition, the probe SB profile of anonymous author is matched with the enrolled SB profiles of known authors. We calculated the similarity scores of each individual feature. Weights have not been applied to the test SB profile to avoid computation overhead during recognition. The similarity scores ($S_{pc}$) of the training and testing profiles are calculated as the total weights $W_t$ of the common items of each feature vector as follows:

$$S_p = \sum_{t \in P_T \cap P_R} W_t$$  \hspace{1cm} (3)

where $P_T$ is an enrolled training feature vector of user $P$ containing a set of weighted items and $P_R$ represents a testing profile containing a set of unweighted items. The similarity score ($S_{pc}$) is normalized as an unit vector to scale its value between 0 to 1. Finally, the normalized similarity scores of retweet (RT), reply (R), hashtag (HT), and domain (D) vectors are fused at score level. Thus, the final similarity score $S_f$ is the average score of all similarity values $S_c$, where $c \in \{RT, R, HT, D\}$.

$$S_f = \frac{\sum S_c}{\sum c}$$  \hspace{1cm} (4)
TABLE I: Style-based Features [21]

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>N_hash_word</td>
<td>Percentage of hashtags</td>
</tr>
<tr>
<td>N_mentions</td>
<td>Percentage of mentions and replies</td>
</tr>
<tr>
<td>N_retweet</td>
<td>Percentage of retweets</td>
</tr>
<tr>
<td>ES_link</td>
<td>Percentage of shared web-links</td>
</tr>
<tr>
<td>N_tweet</td>
<td>Percentage of original tweets</td>
</tr>
<tr>
<td>N_words</td>
<td>Total number of words</td>
</tr>
<tr>
<td>ls_abbreviate</td>
<td>Abbreviation used</td>
</tr>
<tr>
<td>ls_rem_vowels</td>
<td>Summarized form of words used</td>
</tr>
<tr>
<td>N_noun</td>
<td>Number of nouns</td>
</tr>
<tr>
<td>N_verb</td>
<td>Number of verbs</td>
</tr>
<tr>
<td>N_participle</td>
<td>Number of participles</td>
</tr>
<tr>
<td>N_interjection</td>
<td>Number of interjections</td>
</tr>
<tr>
<td>N_pronoun</td>
<td>Number of pronouns</td>
</tr>
<tr>
<td>N_preposition</td>
<td>Number of prepositions</td>
</tr>
<tr>
<td>N_adverb</td>
<td>Number of adverbs</td>
</tr>
<tr>
<td>N_conjunction</td>
<td>Number of conjunctions</td>
</tr>
<tr>
<td>N_special_chars</td>
<td>Number of special characters</td>
</tr>
<tr>
<td>N_capital</td>
<td>Number of capital letters</td>
</tr>
<tr>
<td>Freq_words</td>
<td>Top five frequent words</td>
</tr>
</tbody>
</table>

B. Linguistic Profiles

We implemented two popular methods of linguistic profiles: Bag-Of-Words [1] and style-markers [21]. For both of these methods, we only considered the original tweets written by the users and discarded retweets, replies, and mentions. The two methods are briefly described in the following sub-sections:

1) Bag-Of-Words: Bag-Of-Words is a traditional method of authorship recognition and document classification, which assumes that each document is a collection of independent words [27]. This method considers all content words of documents ignoring grammar or the order of words.

To create a BOW model, at first, we discarded the retweets, replies, and mentions from the latest 200 tweets of each user. Then the original tweets are aggregated to create training sample of each author. We removed web-links, hashtags, stop words, and punctuations from each sample. The words are then tokenized from the writing samples, and BOW corpus has been created considering all words.

2) Style-Markers: Stylometry or style marker is one of the most popular methods for authorship recognition. Unlike BOW, stylometry considers the style of writing, statistics of different words categories, word length, character lengths in words, use of unusual words and symbols, etc. [16]. We implemented a set of stylometric features from a state-of-the-art work [21]. The style-based features used in this work are described in Table I. We used Stanford POS tagger [28] for parts-of-speech tagging of the tweets. Cosine distance has been used to measure the similarity of all style-based features except Freq_words. For Freq_words, we used Jaccard distance. The average similarity scores are then fused to produce the final authorship recognition result.

IV. EXPERIMENTAL RESULTS

A. Data Collection

We collected tweets of 70 users, who produce 100-200 tweets per week on average. An open source social network analyzing tool NodeXL [29] has been used to crawl data from Twitter. Due to Twitter API restrictions, at most 200 recent tweets per user were crawled at a time. We continued our data collection process for four separate sessions over a period of four months, where elapsed time (interval) between two successive sessions was 4-6 weeks. The time intervals allowed us to avoid possible overlapping of data in two successive sessions. Every session contains approximately 200 recent tweets per user. In this article, we denote sessions as S_i, where i = 1, 2, 3, 4. The sessions are numbered according to the sequence of data acquisition. Therefore, session 4 contains the newest 200 tweets of each user, whereas session 1 comprises of the oldest 200 tweets per user. The 200 tweets per user in each session contain text as well as social interactive data such as mentions, replies, retweets, hashtags, and web-links.

B. Experimental Setup

For experimentation, we created six different training and testing set combinations using 4 sessions of data of 70 Twitter users. The six training and testing sessions are - session S_1 vs. S_2, session S_1 vs. S_3, session S_1 vs. S_4, session S_2 vs. S_3, session S_2 vs. S_4, and session S_3 vs. S_4. The different combinations of train and test sets allowed us to evaluate the stability of the features over time. For instance, the S_1 vs. S_4 train and test set combination has at least 10 weeks of time interval and a good recognition performance using S_1 in training and S_4 in testing demonstrates the stability of the features for 10 weeks. In total, 6 sets of experiments were conducted with the 6 combinations of train-test sets to evaluate performance of the proposed SB profile, BOW, and Style markers for authorship recognition. The individual performance of each SB profile features: reply, retweet, hashtag, domains were evaluated as well. All experiments were conducted in closed-set scenario, where the authors were known in the database. Experiments were carried out on Windows 7 operating system, 2.7 GHz Quad-Core Intel Core i7 processor with 16GB RAM. Matlab version R2015a has been used for the implementation and experiments of the proposed method.

C. Results and Discussion

Authorship recognition performance of each of the methods was analyzed by plotting Cumulative Match Characteristics (CMC) curves. CMC curves plot the cumulative probability of obtaining the correct match in the top n positions where n ∈ N. The CMC curves obtained from 6 sets of experiments using SB profile features (Reply, Retweet, Hashtag, and Domain), fused SB profile, BOW, and style markers with session S_1 vs. S_2, session S_1 vs. S_3, session S_1 vs. S_4, session S_2 vs. S_3, session S_2 vs. S_4, and session S_3 vs. S_4 are plotted in Fig. 2a, Fig. 2b, Fig. 2c, Fig. 2d, Fig. 2e, and Fig. 2f, respectively.

From Fig. 2, one can see that the individual interactive features as well as their fused form as SB profile obtained more or less similar recognition performances on all train-test sets. For instance, the rank 1 recognition rate of the fused SB profile on session S_1 vs. S_2, session S_1 vs. S_3, session S_1 vs. S_4, session S_2 vs. S_3, session S_2 vs. S_4, session S_3
Fig. 2: CMC curves of individual SB profile features (Reply, Retweet, Hashtag, and Domain), fused SB profile, BOW, and style markers on a dataset of 70 authors with different training and testing sessions: a) session S₁ vs. S₂, b) session S₁ vs. S₃, c) session S₁ vs. S₄, d) session S₂ vs. S₃, e) session S₂ vs. S₄, f) session S₃ vs. S₄. S₁ vs. S₄ has the longest time interval (at least 10 weeks).

As shown in Fig. 2, the rank-1 recognition rates of Style-based features on session S₁ vs. S₂, session S₁ vs. S₃, session S₁ vs. S₄, session S₂ vs. S₃, session S₂ vs. S₄, session S₃ vs. S₄ are 37%, 31%, 27%, 34%, 31%, 40%, respectively. An interesting trend of style marker is that the rank-1 recognition rate slowly reduced with the increase of time interval. For instance, rank-1 recognition rate of style markers on S₁ vs. S₄ train-test set, which has the maximum time interval, is the lowest (27%). The rank-1 recognition rate of style markers is usually better than BOW (as shown in Fig. 2a-Fig. 2d). BOW obtained better rank-1 recognition rate than style marker on S₂ vs. S₄ and S₃ vs. S₄ (Fig. 2e and Fig. 2f). However, the rank-1 performance of BOW is content dependent since it obtained 38% and 45% rank-1 recognition rate on only two train-test sets - S₂ vs. S₃ and S₃ vs. S₄, respectively. Fig. 2 also shows that both style markers and BOW increase dramatically from
rank-1 to rank-10. Therefore, it is likely to find the correct match within rank-10 using BOW and style marker. In Fig. 2, one can see that SB profile outperformed BOW and style markers on all train-test sets. For all six train-test cases, the recognition performance of SB profile reached to 100% within rank-7 on the dataset.

Three important findings from the experiments are - 1) SB profile outperformed BOW and style-based features regardless of the time intervals, 2) Recognition performance of style markers at rank-1 is usually better than BOW, but the correct match is likely to be present within rank-10 for both BOW and style markers. Whereas, it is highly likely to find the correct match at rank-1 using SB profiles (the rank-1 recognition rate is above 95%), 3) Recognition performance of style markers slowly degrades with the increasing time intervals. Whereas, recognition performance of BOW vary with content rather than time. And, SB profile remains stable over a certain period, where content variation with time has minimal effect.

V. CONCLUSION

In this paper, we presented a new method of micro-blog authorship recognition using social behavioral profile. A comparative analysis between the traditional linguistic profiles and SB profile has been presented as well. Experimental results demonstrated the superiority of the SB profile over linguistic profiles for authorship recognition of tweets. Results also provided evidences that social interactive data in micro-blogs of users contain consistent personal behavioral information. Therefore, ignoring social interactive data during authorship recognition may cause significant loss of information and personality traits. Moreover, SB profiles can be combined with linguistic profiles in order to obtain high recognition rate in Internet-scale (big data) author recognition. The potential applications of the proposed SB profiles include detection of identity theft, deception, and anomaly as well as user authentication and human behavior analysis. Our future works include investigating the differences in recognition performance between retweeted and other SB features on a loosely connected (sparse) social dataset.

ACKNOWLEDGMENT

The authors would like to thank NSERC DISCOVERY program grant RT731064, URG, NSERC ENGAGE, NSERC Vanier CGS, and Alberta Ingenuity for partial support of this project.

REFERENCES