CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. (2016)
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/cpe.3918

SPECIAL ISSUE PAPER

A comparison of classifiers and features for authorship authentication of social networking messages

Jenny S. Li[1], Li-Chiou Chen[1,*,†], John V. Monaco[1], Pranjal Singh[2] and Charles C. Tappert[1]

[1]*Seidenberg School of Computer Science and Information Systems, Pace University, New York City, NY 10038, USA*
[2]*VISA Inc. Technology Center, Bangalore, India*

SUMMARY

This paper develops algorithms and investigates various classifiers to determine the authenticity of short social network postings, an average of 20.6 words, from Facebook. This paper presents and discusses several experiments using a variety of classifiers. The goal of this research is to determine the degree to which such postings can be authenticated as coming from the purported user and not from an intruder. Various sets of stylometry and ad hoc social networking features were developed to categorize 9259 posts from 30 Facebook authors as authentic or non-authentic. An algorithm to utilize machine-learning classifiers for investigating this problem is described, and an additional voting algorithm that combines three classifiers is investigated. This research is one of the first works that focused on authorship authentication in short messages, such as postings on social network sites. The challenges of applying traditional stylometry techniques on short messages are discussed. Experimental results demonstrate an average accuracy rate of 79.6% among 30 users. Further empirical analyses evaluate the effect of sample size, feature selection, user writing style, and classification method on authorship authentication, indicating varying degrees of success compared with previous studies. Copyright © 2016 John Wiley & Sons, Ltd.

Received 2 April 2016; Revised 22 May 2016; Accepted 31 May 2016

KEY WORDS: authorship authentication; social networking; cybersecurity; classification; stylometry; machine learning

1. INTRODUCTION

Social network sites, such as Facebook, Twitter, and Instagram, attract billions of users. These sites provide virtual environments for users to connect with their friends and family or even to make new friends. While users may assume that these sites provide a trusted environment for sharing information, the social network sites could be vulnerable to cyber attacks, resulting in compromised information. For example, hackers could distribute spam messages to users or hack into users' accounts and post fake messages on the users' behalves [1, 2]. It is increasingly important to make sure that a message posted by a user is authentic and not a false message posted by an impostor. *Authorship authentication* is the process of verifying the author legitimacy of a text and could be a means to address the security concerns on online social networks [3].

This research investigates whether it is possible to determine if the acclaimed user actually posted a short message on a social network site by the way the message is written and the posting history of the

*Correspondence to: Li-Chiou Chen, Seidenberg School of Computer Science and Information Systems, Pace University, New York City, NY 10038, USA.
†E-mail: lchen@pace.edu

user. To the best of our knowledge, current social networking sites have not publicly implemented any authorship authentication mechanism. Once a user is logged in, there is no re-authentication or detection of abnormal user behavior. A hacker, after gaining access to a user's account, can act as the user and post messages, comment on the user's circle of friends' posts, or organize events on the user's behalf [4]. The genuine user's friends may not suspect that the posts are not authored by their friend as each message posted is associated with the name of the user. As friends and family leave comments or share the fraudulent posts, the hacker can easily gain information from them.

One of the challenges of this work is the relatively short length of social network messages compared with other sources of text such as e-mails, blogs, or regular articles. Because online social network users can create many posts on a daily basis, it is not practical to apply the same kind of security features used in business transactions, such as *Computers and Humans Apart* (CAPTCHA), to social networking environments. Asking social network users to verify every post they make with such security features would create extra steps for them and reduce the usability of social networks. Hence, there is a need for non-obtrusive authorship authentication procedures for social network postings [5].

The objective of this paper is to investigate how machine learning classification methods can be used to distinguish between authenticate posts and fraudulent posts for a user on a social network site using the historical data maintained by the site. We propose an algorithm for authorship authentication, using a support vector machine (SVM) as the baseline classifier, and evaluate authentication capabilities using a variety of classifiers. In addition, a voting algorithm that utilizes an ensemble of classifiers is investigated. Using a data set of 30 Facebook users' posting history, we evaluate traditional stylometry features originally developed for longer texts and propose a new set of *social networking* features aimed at maximizing authentication accuracy.

Short text such as a post on Facebook makes it challenging in using features and classifiers typically used in authorship authentication. Facebook data collected for this research averaged 20.6 words (or 103 characters) per message, which was a much lower word count per sample comparing to existing literature. Our proposed algorithm has achieved 79.6% average accuracy rate using SVM classifier and comparative results using decision tree classifier and the ensemble of three classifiers.

The rest of this work is organized as follows. Section 2 reviews literature on stylometry and authorship authentication. Section 3 describes our research methodology, data collection, and processing. Section 4 presents the main algorithm used in the proposed model. Section 5 describes our validation procedure and the experimental results. Section 6 contains a discussion about this research, and the conclusions are stated in Section 7.

2. RELATED WORK

2.1. Stylometry

Stylometry refers to the study of linguistic style, typically described by features such as sentence length, word choice, word count, and syntactic structure. Stylometry reflects personal writing styles, defined in terms of stylometric features [6]. There are five common stylometric features [7], including lexical features such as character or word-based, syntactic features such as use of function words ('and', 'but', 'on', etc.) or punctuation, structure features such as how the text is organized, content-specific features such as choice of words within a specific domain, and idiosyncratic features such as misspellings, grammatical mistakes, and deliberate work choices due to, for example, cultural differences.

Stylometry is a well-established means of authorship authentication by using long text such as books and articles. With long text samples, it is typical for researchers to achieve authorship authentication accuracy rates ranging from 70 to over 90%. Baayen *et al*. [8] tested 72 articles produced by eight users, with an average of 908 words per article. They achieved an accuracy of 88.1% by using 50 common function words and eight punctuation symbols. Stamatatos [9] tested 100 messages that ranged from 288 to 812 KB in size (1 KB is roughly 500 words), with a modified *Common N-Gram* method, which achieved an accuracy rate of about 70%. Koppel and Schler [10] used most frequent words to identify the author among 10 authors, with 21 books of varying length

DOI: 10.1002/cpe

per author, and achieved a 95.7% accuracy rate. Iqbal *et al*. [6] used 292 stylometric features, including lexical, syntactic, structural, and topic specific to analyze the Enron Corpus with 158 users, each with 200 e-mails, yielding an 82.9% accuracy rate.

2.2. Authorship identification and authentication

Facebook data collected for this research averaged 20.6 words (or 103 characters) per message, which was a much smaller word count per sample comparing with existing authorship authentication research on long text. Few research had focused on authorship authentication of short text. Table I summarizes research in stylometry that either tested short text or focused on authorship authentication with long text. While the last three papers focused on authorship authentication with long text, the first three papers relied on short text samples but focused on authorship identification (identifying the author of a given text from a list of authors), not authorship authentication (deciding if a given text is written by a specific author). Allison *et al*. [15] used e-mail samples from Enron corpus that were about 500 words per sample. Green *et al*. [12] studied authorship identification using Twitter data, which has a restriction of 140 characters (28 words) per message. Layton *et al*. [13] tested 50 Twitter users, with 120 tweets per user achieving an accuracy rate of 70%.

Zheng *et al*. proposed a framework for authorship identification of online messages, which explored the writing-style features and classification techniques [16]. Most of Zheng's character-based features, word-based features, and syntactic features were included in this study. Zheng's research defined 14 structural features, such as total number of paragraphs and number of lines. These features are not all applicable to social network messages that tend to be short in nature, and only the total number of sentences was used. In addition, Zheng's 11 content-specific features such as 'deal', 'PayPal', and 'obo' were more applicable to messages for online advertisements or other contexts. These content-specific features were not used in this study.

Furthermore, various researchers have investigated the effectiveness of stylometry techniques for authorship authentication using shorter text that ranged from 50 to a few hundreds of words. Their results were not as promising as the results from the long text researches. Angela [17] created an instant message intrusion detection system framework using character frequency analysis to test four users' instant message conversation logs with 69 stylometric features, including sentence structure, predefined specific characters, emoticons, and abbreviations. The naive Bayes classifier yielded the best accuracy, with an average of 68% of data that lie within one standard deviation (SD) of either side of the mean. Corney *et al*. [18] tested four users with 253 e-mails each that ranged from 50–200 words per e-mail. They applied stylistic, structural, and function words as measures and used an SVM [19, 20] as the classification engine, yielding 70.2% identification accuracy.

Table I. Summary of stylometry research in short text.

Ref.	NS	Samples/subject	Sample size	Features	Classifier	AR	P
[11]	9	174–706 e-mails from Enron corpus	75 words average	Word frequency, 2 g, 3 g, and stem words	SVM	86.74%	ID
[12]	12	120–900 tweets from Twitter	140 characters maximum	Hundreds of bag-of-words and 86 style markers	SVM	40.5%	ID
[13]	50	120 tweets from Twitter	140 characters maximum	Character *n*-grams	SCAP	70%	ID
[6]	158	200 e-mails (Enron corpus)	~500 words	292 lexical, syntactic, and structural	Bayesian network	80.60%	AU
[10]	10	21 books	~500 words	250 common words or partial word (*n*-gram)	SVM	95.70%	AU
[14]	30	10 books	10 000 words	228 lexical and syntactic	kNN	91.50%	AU

NS, number of subjects; AR, accuracy rate; P, purpose; ID, authorship identification; AU, authorship authentication.

DOI: 10.1002/cpe

Moreover, Ohtasseb [11] investigated the methods for authorship authentication for online blogs/diaries. The researchers leveraged the Linguistic Inquiry Word Count, MRC Psycholinguistic database, and a collection of syntactic features. They selected 63 Linguistic Inquiry Word Count features and tested eight authors with 301 samples among them to achieve an average of 52.5 to 86% identification accuracy for blog lengths of 100 to 600 words. Alison and Guthrie [11] tested nine users on short e-mails averaging 75 words each, with the number of e-mails per user ranging from 174 to 706, using 2 and 3 g and word frequency measures. Using SVM as the classification engine, they achieved an average of 86.74% accuracy.

3. RESEARCH METHODOLOGY

Our methodology consists of four steps, which include data collection, feature extraction, creation of training data and testing data, and machine learning with a classification method. We first investigated a baseline case using SVM [15,19] as the classifier, while other classifiers, including k-nearest neighbor, naive Bayes, decision tree, and neural network, were used later for comparison. For empirical experiments, we collected data from postings on Facebook, the current largest social network site. This section emphasizes the crucial aspects of the research, including the stylometric system, classification system, feature extraction, and cross validation.

3.1. Stylometric system

As listed in the appendix, we used 233 features in this study, which included 227 stylometric features and six novel social network-specific features. A portion of the 227 stylometric features in this research was selected from a subset of features from Zheng's research [16] to include character-based and word-based features. Other types of feature from Zheng's such as structural features and content-specific features were not used, as they were not applicable to the Facebook data. Zheng *et al*. studied authorship identification of online messages, including messages from e-mail, newsgroup, or chat rooms. Compared with Zheng's studies, Facebook messages are generally shorter than e-mails or newsgroup chats. The average length of our Facebook posts collected was 103 characters, or 20.6 words, assuming an average of five characters per word. The lengths of our Facebook samples are comparable to the length of Twitter messages.

We added six social network-specific features (features 228–233). These features included emoticons (a happy face and a sad face), abbreviations ('LOL'), starting a sentence without an uppercase letter, ending a sentence without a punctuation mark, and not mentioning 'I' or 'We' in the post. These features reflect a more causal writing style in which the user may not care about proper grammar or sentence structure. They write in a colloquial way that is similar to everyday conversation or a style that is commonly seen in chats or short text messages.

3.2. Classification system

For the baseline case, we used SVM Light [21], an implementation of support vector machine [15,19], as the classification engine. SVM Light provides four kernel functions: linear, polynomial, Gaussian radial basis function, and sigmoid tanh. We tested a smaller sample size of 10 users' data with all of the kernel functions. We decided to focus on the default linear kernel function because it produced the best result for earlier tests. Among the 30 users, some were less active than others. They had fewer than 50 posts over a few years. Using half of their data for training and half of it for testing was not practical and may not show a good representation of the users' profiles. To overcome this issue, we used a leave-one-out (LOO) cross-validation method [22] to maximize the number of samples used for training.

Support vector machine Light allows customization of trade-off between training error and margin, which is represented in the C parameter [21]. The value of C provides flexibility to adjust the width of the soft margin from the hyperplane, so that fewer training data fall on the wrong side of the hyperplane. To optimize the value of C, we performed a grid search of C from 0.01 to 2.0, with an

DOI: 10.1002/cpe

incremental of 0.01. $C = 0.8$ produced the best result, which yielded a relatively close false acceptance rate and false rejection rate for our samples.

3.3. Features extractions

Built upon the features identified by others in both long-text and short-text studies, our previous work [23] identified a set of features that include special features commonly used in posts on social networking sites. We created a pre-processing program (implemented in AWK) to extract the 233 features for each Facebook post and generated a feature file for each user. The feature files were passed to the SVM Light [21] program, an implementation of SVM, as inputs for training and classification. For each user, we created an input file that consisted of both positive data from the user and negative data selected randomly from others. These training sets and testing sets for each user were fed into SVM Light for testing. We repeated the process for 30 users. *False acceptance rate* (FAR), *false rejection rate* (FRR), and *accuracy rate* (AR) were calculated for each user's test results:

- FAR = number of false acceptances / number of negative samples * 100%.
- FRR = number of false rejections / number of positive samples * 100%.
- AR = 100% − (FAR + FRR) / 2.

The average result of the 30 users was calculated for each test case. The highest accuracy rate for each test case was also recorded. We repeated the process for all 12 test cases. We also performed three runs for each test case to obtain a more representative result. With the LOO method (discussed in Section 3.4), we ran more than 666 000 tests for 12 test cases. Refer to Section 4 for test results.

3.4. Cross validation

A LOO cross-validation procedure is used to obtain accuracy results. In each fold, one sample from a single user is designated as the testing set, and the remainder of the data is used for training. This process is repeated so that each sample is designated as the testing set once. Assuming that there are N positive samples and N negative samples for each user, this process will generate $2N$ -1 predictions, excluding one sample as a test case. The average result of the two N-1 predictions becomes the result of the user for this test case on the first run. The operations in the preceding texts are repeatable for rest of the data. The AR, FAR, and FRR are calculated for this user. The average result from all of the users is calculated, including AR, FAR, FRR, and SD. This completes one run of the test case for all users.

To achieve a more representative result, the process in the preceding texts is repeated thrice using a random selection of positive and negative data for each user. The average result (AR, FAR and FRR, and SD) of all of the three runs is then calculated. This completes one test case.

4. ALGORITHM

This section summarizes two algorithms in which the first one extracts features from Facebook postings collected and the second one trains the classifiers and obtains classification results. A list of notations is mentioned in the succeeding texts.

- N: the number of users. In our case, we collected data from 30 users. $N = 30$.
- M_i: the number of posts from user i. i = 1 to N.
- INFILE_i: the ith input file with posts from user i collected from Facebook.

The pre-processing algorithm parsed Facebook posts for various features. The algorithm calculated the value for each feature tested in an experiment and stored the values and feature identifier in an output files for the training and validation of the classifiers. The number of features generated was based on the test cases in the experimental design. Section 5.2 will describe the 12 test cases in our research. E refers to the number of features in a test case.

DOI: 10.1002/cpe

Pre-processing algorithm

```
For i = 1 to N {
  R = Read a post from INFILE_i.
  While (R ≠ the end of the file)    {
      For k = 1 to E {
                  Calculate the un-normalized feature value V_k from R based
                  on Table A-I in the <Appendix.
                  Store the identifier and the value of the feature, (k, V_k), in
                  the output file OUTFILE_i.
      }        R = Read a post from INFILE_i.
  }
}
```

The training and validation algorithm took the output files, OUTFILE_i, from the pre-processing algorithm to generate validation data set T_i. LOO validation was used to validate each test instance in the data set, which consisted of the equal number of positive and negative instances. The average results were calculated at the end when instances from all users were tested. We run the same algorithm multiple times for each user to reduce the bias that may occur when random negative samples were drawn from other users.

Training and validation algorithm

```
For i = 1 to N      {
    Create training and validation data set, T_i, for user_i.
        T_i = OUTFILE_i (positive data set) + the equal number of samples
        randomly selected from OUTFILE_j, where j = (1 to N) and j ≠ i (negative
        data set).
    For k = 1 to M_i {
            Train the classifier C_i with all instances in T_i, excluding instance k.
            Test instance k using LOO validation on the trained classifier C_i.
                Output (user identifier, i; case label, 0 for positive sample and 1 for
                negative sample; test result, 0 for positive prediction and 1 for
                negative prediction).
    }       Calculate and output (AR_i, FAR_i, and FRR_i).
}  Calculate the average results, AR, FAR, and FRR, and SD from N users.
```

5. DATA COLLECTION AND EXPERIMENT RESULTS

This section described our process of data collection, the experimental configurations, and the analysis to derive implications from our results.

5.1. Data collection and feature extraction

We collected Facebook posts from 30 users, including six friends who do not know each other and agreed to provide their posts and 24 public figures (such as movie stars, athletes, journalists, politicians, etc.) that have their posts publicly accessible. To guarantee the confidentiality of these users, their identities would not be disclosed. Their data remained anonymous throughout the study. We set a goal to collect at least 400 posts from each user, which represented a few years of postings. However, some users did not post as often, and they ended up with fewer than 400 posts. As a result, we collected a total of 9259 posts over the last 4 years among 30 users. The average number of posts per user was 308.6. The average length of these posts was approximately 20.6 words, assuming five characters per word. Table II in the succeeding texts provides a summary of the data collected from the 30 users.

DOI: 10.1002/cpe

Table II. A summary of Facebook data collected.

	Number of posts	Number of words	Average words per post
User 1	400	7970	19.9
User 2	400	5947	14.9
User 3	400	8908	22.3
User 4	400	8908	22.3
User 5	400	6733	18.8
User 6	400	8215	20.5
User 7	400	10 843	27.1
User 8	400	6681	16.7
User 9	400	6936	17.3
User 10	400	3440	8.6
User 11	400	10 218	25.5
User 12	400	2672	6.7
User 13	374	11 699	31.3
User 14	338	13 643	40.4
User 15	330	6479	19.6
User 16	281	5600	19.9
User 17	205	3887	19.0
User 18	109	2880	26.4
User 19	106	1219	11.5
User 20	45	444	9.9
User 21	400	3607	9.0
User 22	400	18 830	47.1
User 23	360	5490	15.3
User 24	300	6312	21.0
User 25	300	5080	16.9
User 26	283	6765	23.9
User 27	242	5434	22.5
User 28	230	3175	13.8
User 29	108	1710	15.8
User 30	48	744	15.5
Total	9259	190 469	20.6

The data collected were static posts, which were initiated by the users and were not responses from these users to others. The responses to posts are typically private information that is not accessible by others or public. Because of the scope of this research, we collected only textual inputs and do not collect metadata, such as the date of postings, geographical location of the user, or applications being used for the postings.

5.2. Impact of features

We conducted 12 sets of tests on Facebook data using the 233 selected features and SVM as the classifier. These tests aimed to discover the performance of stylometric and social network-specific features, combined or separated, for authorship authentication.

In Table III, the combined use of stylometric features and social network-specific features (Test 1) produced the best accuracy rate of 79.6% and lowest SD among the tests. Stylometric features by themselves (Test 2) yielded a 78.9% accuracy rate, almost as good as the combination of stylometry and social network-specific features together. This showed the selected six social network-specific features provided a slight improvement to the overall accuracy when being combined with stylometric features.

In general, the six social special features alone (Test 3) were not as reliable. The list only yielded a 69.8% accuracy rate for the 30 users on average. However, we observed a phenomenon that social network-specific features could be helpful in determining authorship if a user frequently used some of these features such as emoticons or others. The highest accuracy rate found among the 30 users for the six social network-specific features was 96.6 versus 95.3% when combined with stylometric

DOI: 10.1002/cpe

Table III. Authorship authentication tests with 233 features on 30 users' Facebook data.

Test	Features	AR	FAR	FRR	HAR	SD
Test 1	All features (233 features)	79.6	14.3	26.5	95.2	6.7
Test 2	Stylometry only (227 features)	78.9	15.1	27.0	94.9	7.7
Test 3	Social network-specific (6 features)	69.8	24.3	36.0	96.6	12.8
Test 4	Char-based (50 features)	76.0	17.7	30.4	98.4	8.8
Test 5	Punctuations (8 features)	73.8	26.6	25.8	98.3	12.6
Test 6	Function words (150 features)	72.9	22.3	31.9	96.3	9.9
Test 7	No. of sentences (1 feature)	53.6	50.8	42.1	87.5	16.7
Test 8	Word-based (8 features)	74.1	21.3	30.6	95.5	10.0
Test 9	Popular function words (33 features)	71.6	24.9	32.0	96.6	10.6
Test 10	Smilies (2 features)	67.8	54.4	10.0	99.6	18.1
Test 11	Missing upper case, period, etc.(2 features)	67.9	28.2	36.0	98.6	15.9
Test 12	Not mentioning 'I' and 'We' (1 feature)	60.8	40.5	38.0	98.0	18.4

AR, accuracy rate; HAR, highest accuracy rate; SD, standard deviation.

features as in Test 1. This result was further supported by tests on individual social network-specific features. The highest accuracy rate found among the 30 users using only the smiley feature (Test 10) was 99.6%. This demonstrated that one user used smilies on more than 82% of his or her posts, while others rarely used smilies. Hence, this user's writing style was more distinctive. In addition, the SD of the accuracy rate is among the highest for Test 10, which indicates a wide range of writing styles among the users, particularly the use of smilies.

The results for Test 11 indicate that beginning of sentence capitalization varies across users. A manual inspection of the data reveals some instances in which beginning of sentence capitalization is not used. These include starting a sentence with a hashtag by using the '#' sign or tag a person by using the '@' character or starting a sentence with a quote from other people by using a quotation mark. Similarly, we observed many instances where proper punctuation was not used to end a post. These include the use of many punctuations such as '!!!' or '???', use of different symbols to represent an emoticon or a face such as '> <' or ':−/', use of character combinations to express feelings such as 'xoxo' for hugs and kisses or '<3' for love, use of a signature such as '-xxx' where 'xxx' is the user's name or signature, or incomplete sentences as caption for a picture or link. When a user wrote with proper capitalization for opening a sentence and used proper punctuation to close a sentence, the user's writing style would stand out from the rest. This was reflected with the highest accuracy rate of 98.6% from the result of Test 11.

Test 12 investigated how often users talked about themselves by using 'I' or 'We' versus other topics. It was a surprise that a majority of our Facebook users talked about other topics for more than 60% in their posts. For users who talked about themselves most of the time, these users' writing styles were more distinctive. The highest accuracy rate was 98% for one of these users as he or she talked about himself or herself for 88% of the time.

Furthermore, most Facebook users in our study posted very short messages comparable with the messages posted on Twitter, which has a stricter limitation on the length of the messages. The average number of words per post was 20.6. Stylometric features that are character based (Test 4) would be more effective to determine authorship than word-based features (Test 8). There may not be enough words for word-based features to take effect for developing a user writing style profile. Among different types of stylometric feature, character-based stylometric features (Test 4) alone yielded best accuracy rate of 76%, followed by word-based features (Test 8), with a 74.1% accuracy rate, punctuation-based features (Test 5) that yielded a 73.8% accuracy rate, and function words (Test 6) that generated a 72.9% accuracy rate. Sentence-based features (Test 7) showed the worst performance, with a 53.6% accuracy rate.

Because users tended to write short messages, most users would end up with one or two sentences. There was little to differentiate among messages by just looking at the number of sentences. The 150 function words were used in Test 6, including 'about', 'from', 'if', 'and', 'but' etc. As our Facebook messages were short, it is most likely that only a small subset of the function words was used in each post. We further selected a subset of 33 popular function words in Test 9. These words were

DOI: 10.1002/cpe

used more than 10% of the total number of posts collected. The short list of function words yielded an accuracy rate of 71.6%, while the full list of function words yielded 72.9% (in Test 6). More features being used would not harm the result of the tests. However, it costs more computational time to calculate values of more features. It would be a design decision of the social network authorship authentication provider to decide on the trade-offs between computational effort and accuracy.

5.3. *Impact of number of users*

We used all 233 features to test a batch of 10, 20, and 30 users for the impact of number of users used (Table V). Compare 10 with 20 users, testing 10 users yielded an 81.6% accuracy rate, while 20 users (that included the same 10 users as before) yielded a 79.8% accuracy rate. The tests of 30 users (that included the previous 20 users) yielded a 79.6% accuracy rate. There was a slight advantage of using only 10 users as the accuracy rate was slightly better than 20 or 30 users. However, the results between 20 and 30 users were too close to conclude that the increase in number of users being tested decreased the accuracy rate in authorship authentication.

Users with distinctive writing styles were easier to be differentiated from the rest. In our tests (Table IV), we separated the users into three groups of 10 users. Both second and third groups contain users with very distinctive writing styles, as shown by the highest accuracy rates of 95.3 and 94.9% for the second group and the third group, respectively. The highest accuracy rate from the first group was only 85.4%. Therefore, a larger group of users with more distinctive writing styles can out-perform a smaller group of users with less distinctive writing styles (Table V).

5.4. *Impact of number of features*

Would more features clutter the analysis or decrease the accuracy rate? Our results from Table III showed a tendency that tests with more features showed a higher accuracy rate and a smaller SD. Test 7 (number of sentences) and Test 12 (missing 'I' and 'We') both had one feature. They produced the two worst results in terms of low accuracy rate and large SD. Test 1 with all 233 features produced the best accuracy rate and had the smallest SD. So far, all of the test results supported the argument that testing with more features would produce a better result except for the case for Test 4. Test 4 with 50 character-based features produced a better result (76% accuracy rate and 8.8% SD) than Test 6 with 150 function words (72.9% accuracy rate and 9.9% SD). This simply confirmed that character-based features were more desirable measures for short-text authorship authentication than word-based features.

Table IV. Testing different user groups with 23 features.

Test ID	Group of users	AR	FAR	FRR	HAR
A	Group A: 1–10	77.9	21.7	22.4	85.4
B	Group B: 11–20	81.6	9.8	26.9	95.3
C	Group C: 21–30	79.3	11.4	30.1	94.9

AR, accuracy rate; HAR, highest accuracy rate.

Table V. Testing different sizes of user groups with 23 features.

Test ID	Number of users	AR	FAR	FRR	HAR
Test 1a	10	77.9	21.7	22.4	85.4
Test 1b	20	79.8	15.8	24.6	95.3
Test 1c	30	79.6	14.3	26.5	95.2

AR, accuracy rate; HAR, highest accuracy rate.

DOI: 10.1002/cpe

5.5. Testing with normalized stylometric features

The list of selected 233 combined stylometric features and social network-specific features described in the preceding texts was not normalized. The frequency of each feature as it appeared in each post was counted. Stylometric features that were used in Monaco *et al.*'s research [14] were used for re-testing. The goal was to investigate if another set of stylometric features would yield better results on our Facebook data. Monaco's features were normalized, which represented the ratio of each feature against the whole message. Monaco used his set of 228 stylometric features for authorship identification of 30 book authors. Each had 10 book samples; each book was about 10 000 words or longer.

Using Facebook data, Monaco's stylometric features, and SVM, the test result was 77% accurate, with 17.4 FAR and 28.7 FRR. This result was very close to the 227 stylometric feature test from Test 2 (Table III Test 2) that yielded 78.9% accuracy with 15.1% FAR and 27% FRR. Even though Monaco's features were not exactly the same as our features, there were duplications, including the character-based, word-based, and some syntax-based features. It was hard to conclude if normalized or un-normalized features were more applicable for testing with short text. Both results were similar, although un-normalized features showed slight improvement in performance for short messages collected in our study.

5.6. Testing with the kNN algorithm for classification

To compare with another classification algorithm, the data were re-tested using the k-nearest neighbor algorithm. It uses Euclidean distance to classify the unknown difference vectors, with a reference set composed of the differences between all combinations of the claimed user's enrolled vector (within-person) and the differences between the claimed user and every other user (between-person). The differences of difference vectors are being calculated [14].

Using the k-nearest neighbor method, the average accuracy among the 30 users for Test 1 (with all 233 features) was 65.5%. The result showed that SVM yielded a much better result of 79.6% accuracy rate (Table III Test 1) than the k-nearest neighbor algorithm for short messages.

5.7. Testing with more classification methods

Different classification methods, including SVM in linear, RBF, and polynomial kernel functions, naive Bayes, decision tree, and neural network (NN), were tested with the first 10 users' data, each of which had 400 samples. The objective was to determine which classification method works better for authorship authentication with short messages. Can performance be improved from the result of SVM? All tests were performed on the first 10 users' data with 233 stylometric features and the leave-one-out method using Weka [24]. Unlike SVM Light that only uses SVM as the classification method, Weka is a machine learning program that offers a variety of classification algorithms.

Seven test cases with different classification methods were performed to test their performance using the same training and testing data. Among all of the tests in Table VI, decision tree performed the best,

Table VI. Testing different classification methods.

	CTest1, SVM, linear	CTest2, SVM, RBF	CTest3, SVM, poly	CTest4, Naïve Bayes	CTest5, decision tree	CTest6, NN 2 layers	CTest7, NN 3 layers
User 1	72.8	74.5	74.8	69.0	77.5	71.3	67.1
User 2	82.7	74.3	79.5	66.0	80.9	80.2	78.2
User 3	63.7	59.3	65.0	63.0	68.3	56.6	52.2
User 4	79.1	65.1	78.1	55.4	73.3	74.3	56.6
User 5	84.3	73.2	79.8	70.6	83.6	81.6	81.3
User 6	77.5	70.7	73.2	63.8	74.3	71.0	69.1
User 7	82.5	68.7	80.1	73.9	79.6	76.6	59.7
User 8	71.9	67.3	73.6	63.4	79.3	69.3	67.0
User 9	77.1	70.2	81.0	69.8	82.9	70.5	70.3
User 10	76.1	73.0	75.6	71.9	74.6	72.1	66.5
AVG (SD)	76.8 (6.2)	69.6 (4.8)	76.1 (4.8)	66.6 (5.5)	77.4 (4.8)	72.3 (6.9)	66.8 (8.5)

AVG, average accuracy rate of the 10 users; SD, standard deviation.

DOI: 10.1002/cpe

with a 77.4% accuracy rate (in CTest5), followed by SVM linear kernel function, with a 76.8% accuracy rate (in CTest1). Naive Bayes (in CTest4) and neural network (in CTest7) with three-layer analysis performed the worst. SVM linear kernel function (in CTest1) and SVM polynomial function (in CTest2) generate similar results. But SVM linear function was recommended over polynomial function, as polynomial function took more computing effort, which resulted in much longer time of completion. Comparing the two neural network tests, the two-layer test (in CTest6) that generated an accuracy rate of 72.3% performed better than the three-layer test (in CTest7) that generated an accuracy rate of 66.8%. Increasing from two-layer to three-layer analysis, the neural network method declined in performance with the given samples. To sum up these test results, both decision tree and SVM linear kernel function showed good performance over the other methods, with authorship authentication using short messages.

5.8. Testing with a combination of classifiers

Decision tree and SVM linear kernel yielded comparable results, although decision tree had a slightly better average performance with a lower SD. However, when we compared the accuracy rate of these two classifiers for each of the 10 users tested in the previous subsection, the best test result alternated between the two classifiers. We suspected that we would be able to improve our results by leveraging a combination of classifiers.

We developed a voting algorithm to determine the decision of a test based on the majority votes of three classifiers. Decision tree, SVM linear, and neural network were selected for our tests because the first two classifiers performed better in our previous tests than the rest, and neural network was added as the third classifier to break the tie when decision tree and SVM produced different results.

Voting algorithm

C1 = decision tree, C2 = SVM, and C3 = neural network. $N = 10$ in our experiments since we only tested the first 10 users who we have collected 400 samples each. Therefore, the sample size for each user, $M = 400$.

```
For i = 1 to N {
    Create training and validation data set, T_i, for user_i.
        T_i = OUTFILE_i (positive data set) + the equal number of samples
        randomly selected from OUTFILE_j where j = (1 to N) and j ≠ i (negative
        data set).
      For k = 1 to M {
            For q = 1 to 3{
                  Train the classifier C_iq with data in T_i, excluding instance k.
                  Test instance k using LOO validation on the trained classifier C_iq.
                  Output (classifier, q; user identifier, i; case label, 1 for positive
                  sample and 0 for negative sample; test result, 1 for positive
                  prediction and 0 for negative prediction).
            } // end of q
            S = sum of the predictions from the three classifiers
            If (S>=2)
                  Output (user identifier, i; case label, 1 for positive sample and 0
                  for negative sample; test result = 1)
            else
                  Output (user identifier, i; case label, 1 for positive sample and 0
                  for negative sample; test result = 0)
    } Calculate and output (AR_i, FAR_i, and FRR_i).
}     Calculate the average results, AR, FAR, and FRR, and SD from N users.
```

The test results (Table VII) showed that decision tree, SVM, and the voting algorithm yielded similar performance, while decision tree (78% accuracy) showed slight improvement over SVM (76.9%

DOI: 10.1002/cpe

Table VII. Test result of three classifiers and the voting algorithm.

	SVM, linear	Decision tree	Neural network, 2 Layers	Voting algorithm
User 1	72.8	78.1	71.3	74.8
User 2	82.9	80.9	78.2	82.7
User 3	63.7	68.3	59.5	66.8
User 4	79.2	75.2	75.3	79.0
User 5	84.1	85.2	61.0	83.6
User 6	78.4	75.3	63.0	76.0
User 7	82.5	78.7	76.9	82.4
User 8	72.2	78.7	63.3	72.8
User 9	77.4	84.2	68.8	79.1
User 10	76.3	75.7	71.7	75.2
AVG (SD)	76.9 (6.2)	78 (4.9)	68.9 (6.8)	77.2 (5.2)

AVG, average accuracy rate of the 10 users; SD, standard deviation of accuracy rate.

accuracy) and the voting algorithm (77.2% accuracy) and a substantial improvement over neural network (68.9% accuracy). Although the voting algorithm showed improvement over SVM, it did not outperform decision tree. Future work could investigate more sophisticated classifier fusion algorithms.

6. DISCUSSIONS

Although authorship authentication using stylometry has been a well-known area, research has mostly been performed in long text. This research was one of the first research that attempted to use Facebook data to study authorship authentication with short text. While online social networks have gained tremendous popularity, they have also created security threats to users [25], and many prior researches addressed the optimization issue of cloud solutions [26]. Our research aims at providing an alternative solution to address the threats.

We faced two challenges. The first challenge was that social networking messages tend to be much shorter than novels, blogs, or e-mails, and our concern was whether traditional stylometric features would be effective in authorship authentication for these short messages. The second challenge was that the limited number of posts from some users made it hard to learn the authors' writing styles from existing posts. Because some users post infrequently, it is not reasonable to divide these users' data into halves for training and testing. Because of these challenges, the feature sets used have to be adjusted to the nature of the data.

Unlike Twitter, which has mandated that its users write short messages with a maximum of 140 characters, Facebook's users have the freedom to write as little or as much as they wish, up to 60 000 characters [16]. The result of the data collected for this research showed that users still preferred to keep their messages short on Facebook. The Facebook posts collected had an average of 20.6 words per post, which was much shorter than blogs or e-mails, typical data used in the previous authorship authentication study. The data we collected are comparable to the data in other Facebook study [27].

The test results of using all 233 combined stylometric and social network-specific features showed an accuracy rate of 79.6% when verifying whether a message was written by the real user. As we pointed out in Section 2.2 (Table I), our study is unique because none of the studies we have found at this point specifically focuses on authorship authentication with short texts such as Facebook posts. The closest study to ours is the Twitter study for authorship identification [13], which has an accuracy rate of 70% with a much lower sample size comparing to ours.

Our results provided a prediction to questions such as 'Does this message look like a message that would be written by the user?' and 'Can we trust this message?' We tested individual subsets of our features and showed the impact of different feature sets. Our study gives social network providers an overview of the trade-offs in case they are interested in building an authorship authentication solution to protect their users' accounts. In the future, this research can be extended to use the same list of features for testing long messages such as blog, e-mails, novels, etc. By doing so, we can evaluate the effectiveness of these features in long messages versus short messages for authorship authentication. Idiosyncratic features such as misspellings and grammatical mistakes could be

DOI: 10.1002/cpe

applicable to social network postings for some users. This research can be extended to include these features. However, a more sophisticated feature extraction program is needed to determine whether any of these features exist in the postings collected.

A number of distinct classifiers were tested on the same Facebook data. The results showed that SVM and decision tree were effective when used for authorship authentication that used short messages. SVM had a significant improvement in accuracy over kNN when used with the same feature set on Facebook data, and as a result, kNN is not recommended for use on short text. In addition, we investigated other classification methods using the short messages from the first 10 users of our data set. We discovered that decision tree performed the best on short messages, although the results were still comparable to SVM linear and the voting algorithm that we developed. Further investigation with a larger data set is needed to validate the results.

The usage of the six social network-specific features in this research proved that users who adopted a social writing style are more distinguishable than others. This research served as a proof of concept for improving authorship authentication in social network sites. Currently, state-of-the-art social network sites, including Facebook, only use textual login and password to authenticate their users. To authenticate users with the messages they write could be used as an additional authentication mechanism.

The performance of our approach should be evaluated in three different parts: (1) pre-processing of the messages to extract features, (2) classifier training, and (3) validation. First of all, the performance of the pre-processing algorithm depends on the length of the message (M) and the number of features (F). The complexity of the algorithm is $O(M * F)$. Because our approach focuses in short text, an average of 20.6 words in our case, M is negligible. Secondly, the performance of classifier training depends on the complexity of the classifier that is used and the number of instances used in training. Among the classifiers that we used for training, decision tree had the best performance, SVM linear was the second, and neural network (two layers) took the longest among the three classifiers, and the other classifiers took even longer. The performance can also be improved using cluster-computing platform, such as Spark [28]. Thirdly, the time took for validating a test instance for a trained classifier is negligible because the complexity is $O(1)$. Finally, when deploying our approach for authorship authentication on social networking sites, the classifier should have already been trained using historical data. That is, the first two parts would have been carried out prior to making a prediction for an authentication application in real time. Therefore, the complexity in this case is equal to the complexity of testing one instance, $O(1)$, plus the complexity of extracting features from the instance, $O(F)$.

Several future work can be conducted based our research. First, this research only leveraged three types of stylometric feature: lexical, syntactic, and structural. Other types of stylometric feature, including content-based or idiosyncratic features, could potentially be explored. Secondly, our approach can be further validated using data from other social network sites or applications that tend to generate short text. Thirdly, to reduce any instances of users copying each other in their writing styles, we selected users who do not know each other and are from a diverse group. It would be interesting to see a study on authorship authentication with Facebook data from a group of friends who interact with each other on a regular basis. They may imitate each other's behavior or writing style, including use of emoticons, hashtags, links, or others. In that case, additional features that include metadata of postings might be needed to distinguish their writing styles. Finally, further investigation into the combination of multiple classifiers, such as our voting algorithm, could be a promising step to improve the current results.

7. CONCLUSIONS

This paper developed algorithms and investigated various classifiers to determine the authenticity of short social network postings. This research is one of the first works that focused on authorship authentication in short text, which is a more challenging research problem than the same problem in long text due to the length of each sample. Through empirical study, this research aimed to identify method that can best determine whether a new and possibly disputed message can be authored by the same user when a set of messages posed by a user are given. Various classification methods were investigated, and a voting algorithm of combining classifiers was developed. Based on a data set of 9259 social network posts from 30 users, our experimental evaluations had shown that an accuracy rate of 79.6% was achieved

DOI: 10.1002/cpe

when verifying if a message was written by the same user. SVM with linear kernel and decision tree yielded the best performance in addressing our research problem. The voting algorithm that we developed also produced comparable results between SVM and decision tree.

APPENDIX

TABLE A-I. FEATURES USED IN THIS STUDY.

Feature	Description of the feature
Character-based	
Feature 1	Number of characters
Feature 2	Number of alphabets
Feature 3	Number of uppercase characters
Feature 4–29	Number of alphabet a–z
Feature 30–50	Number of special character '~ @ # $ % ^ & * - = + > < [ ] { } / \\ |'
Syntactic-based	
Feature 51–58	Number of punctuation ', . ? ! : ; \" ' '
Feature 59–208	Function words including 'a, about, above, after, all, although, am, among, an, and, another, any, anybody, anyone, anything, are, around, as, at, be, because, before, behind, below, beside, between, both, but, by, can, cos, do, down, each, either, enough, every, everybody, everyone, everything, few, following, for, from, have, he, her, him, I, if, in, including, inside, into, is, it, its, latter, less, like, little, lots, many, me, more, most, much, my, need, neither, no, nobody, none, nor, nothing, of, off, on, once, one, onto, opposite, or, our, outside, over, own, past, per, plenty, plus, regarding, same, several, she, should, since, so, some, somebody, someone, something, such, than, that, he, their, them, these, they, this, those, though, through, till, to, toward, towards, under, unless, unlike, until, up, upon, us, used, via, we, what, whatever, when, where, whether, which, while, who, whoever, whom, whose, will, with, within, without, worth, would, yes, you, your'
Feature 209	Total number of sentences
Word-based	
Feature 210	Total number of words
Feature 211	Total number of short words (less than four characters)
Feature 212	Average word length
Feature 213	Average sentence length in terms of character
Feature 214	Average sentence length in terms of word
Feature 215	Number of words with 1 char
Feature 216	Number of words with 2 chars
Feature 217	Number of words with 3 chars
Feature 218	Number of words with 4 chars
Feature 219	Number of words with 5 chars
Feature 220	Number of words with 6 chars
Feature 221	Number of words with 7 chars
Feature 222	Number of words with 8 chars
Feature 223	Number of words with 9 chars
Feature 224	Number of words with 10 chars
Feature 225	Number of words with 11 chars
Features 226	Number of words with 12 chars
Features 227	Number of words with more than 12 chars
Social networking-based	
Feature 228	Frequency of a smiley face ':)'
Feature 229	Frequency of a sad face ':('
Feature 230	Frequency of 'LOL'
Feature 231	Frequency of missing an uppercase letter when starting a sentence
Feature 232	Frequency of missing a period or other punctuation to end a sentence
Feature 233	Frequency of missing the word 'I' or 'We' when starting a sentence

ACKNOWLEDGEMENT

The authors would like to acknowledge the support from the National Science Foundation under grant no. 1241585. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the US government.

DOI: 10.1002/cpe

REFERENCES

1. BBC. Facebook protects users following adobe hack attack, 2013. (Available from: url=http://www.bbc.com/news/technology-24925874) [Accessed on 13 November 2013].
2. Tao L, Golikov S, Gai K, Qiu M. A reusable software component for integrated syntax and semantic validation for services computing. In *9th IEEE International Symposium on Service-Oriented System Engineering*, pages 127–132, San Francisco Bay, USA, 2015.
3. Li Y, Dai W, Ming Z, Qiu M. Privacy protection for preventing data over-collection in smart city. *IEEE Transactions on Computers* 2015; **PP**:1.
4. Qiu M, Gai K, Thuraisingham B, Tao L, Zhao H. Proactive user-centric secure data scheme using attribute-based semantic access controls for mobile clouds in financial industry. *Future Generation Computer Systems* 2016; **PP**(1).
5. Gai K, Qiu M, Thuraisingham B, Tao L. Proactive attribute-based secure data schema for mobile cloud in financial industry. In *The IEEE International Symposium on Big Data Security on Cloud*; *17th IEEE International Conference on High Performance Computing and Communications*, pages 1332–1337, New York, USA, 2015.
6. Iqbal F, Khan L, Fung B, Debbabi M. E-mail authorship verification for forensic investigation. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, pages 1591–1598, Western Switzerland, Sierre, 2010. ACM.
7. Abbasi A, Chen H. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems* 2005; **20**(5):67–75.
8. Baayen H, Halteren H, Neijt A, Tweedie F. An experiment in authorship attribution. In *6th JADT*, pages 29–37, 2002.
9. Stamatatos E. Author identification using imbalanced and limited training texts. In *18th International Workshop on Database and Expert Systems Applications*, pages 237–241, Regensburg, 2007. IEEE.
10. Koppel M, Schler J. Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning*, page 62, Banff, Alberta, Canada, 2004. ACM.
11. Allison B, Guthrie L. *Authorship Attribution of E-mail: Comparing Classifiers over a New Corpus for Evaluation*. In *LREC*: Morocco, 2008.
12. Green R, Sheppard J. Comparing frequency- and style-based features for Twitter author identification, Proc. Twenty-Sixth International Florida Artificial Intelligence Research Society Conference, 2013.
13. Layton R, Watters P, Dazeley R. Authorship attribution for twitter in 140 characters or less. In *The Second Cybercrime and Trustworthy Computing Workshop*, pages 1–8, Ballarat, Victoria, Australia, 2010. IEEE.
14. Monaco J, Stewart J, Cha S, Tappert C. Behavioral biometric verification of student identity in online course assessment and authentication of authors in literary works. In *Sixth International Conference on Biometrics*: *Theory, Applications and Systems*, pages 1–8, Arlington, VA, 2013. IEEE.
15. Cortes C, Vapnik V. Machine learning. *Support Vector Networks* 1995:273–279.
16. Zheng R, Li J, Chen H, Huang Z. A framework for authorship identification of online messages: writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology* 2006; **57**(3):378–393.
17. Orebaugh A. An instant messaging intrusion detection system framework: using character frequency analysis for authorship identification and validation. In *Proceedings 2006 40th Annual IEEE International Carnahan Conferences Security Technology*, pages 160–172, Lexington, KY, 2006. IEEE.
18. Corney M, De Vel O, Anderson A, Mohay G. Gender-preferential text mining of e-mail discourse. In *The 18th Annual Computer Security Applications Conference*, pages 282–289, Las Vegas, NV, USA, 2002. IEEE.
19. Vapnik V. *The Nature of Statistical Learning Theory*. Springer Science & Business Media, Springer-Verlag, New York, USA, 2013.
20. Yin H, Gai K. An empirical study on preprocessing high-dimensional class-imbalanced data for classification. In *The IEEE International Symposium on Big Data Security on Cloud*; *17th IEEE International Conference on High Performance Computing and Communications*, pages 1314–1319, New York, USA, 2015.
21. Joachims T. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Kluwer Academic Publishers Norwell, MA, USA 2002.
22. Kearns M, Ron D. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation* 1999; **11**(6):1427–1453.
23. Li, JS, Monaco JV, Chen L-C, Tappert CC. Authorship authentication using short messages from social networking sites, 2014 IEEE 11th International Conference on e-Business Engineering, 2014.
24. Witten IH, Frank E, Hall MA. *Data Mining: Practical Machine Learning Tools and Techniques* (3 edn). Morgan Kaufmann, Burlington, MA, USA 2011.
25. Luo W, Liu J, Liu J, Fan C. An analysis of security in social networks. In *DASC'09. Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing*, pages 648–651, Chengdu, China, 2009. IEEE.
26. Qiu M, Ming Z, Li J, Gai K, Zong Z. Phase-change memory optimization for green cloud with genetic algorithm. *IEEE Transactions on Computers* 2015; **64**(12):3528–3540.
27. Hussain A, Vatrapu R, Hardt D, Jaffari Z. Social data analytics tool: a demonstrative case study of methodology and software. In *Analyzing Social Media Data and Web Networks*. Palgrave Macmillan, UK 2014; 99–118.
28. Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I. Spark: cluster computing with working sets. *HotCloud* 2010; **10**:10–10.

DOI: 10.1002/cpe