An Investigation of Authorship Authentication in Short Messages from a Social Networking Site

by
Jenny S. Li

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Professional Studies
in Computing

at

School of Computer Science and Information Systems

Pace University

May 2015

We hereby certify that this dissertation, submitted by XXX XXXX, satisfies the dissertation requirements for the degree of *Doctor of Professional Studies in Computing* and has been approved.

_____-_____

Dr. Li-Chiou Chen Date
Chairperson of Dissertation Committee

_____-_____

Dr. Charles Tappert Date
Dissertation Committee Member

_____-_____

Dr. Fred Grossman Date
Dissertation Committee Member

School of Computer Science and Information Systems
Pace University 2015

Abstract

An Investigation of Authorship Authentication in Short Text from a Social Networking Site

by
Jenny S. Li

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Professional Studies
in Computing

May 2015

An authorship authentication system was presented in this research to assess authorship authentication in short text that was extracted from sample posts of a social networking site. Authorship authentication is one of the trending security problems for social networking sites. Stylometry has been a well-known technology for authenticating an author to a disputed document in question. Authorship authentication in short text from social networking sites is a relatively newer domain to be explored. The goal of this research is to determine the degree to which social networking postings can be authenticated as coming from the purported user and not from an intruder. Facebook data was used for illustration.

The proposed research solution is an authorship authentication system that includes the use of 233 features (a combination of 227 stylometric features and 6 social network specific features), Support Vector Machine (SVM) Linear kernel function, and the Leave-One-Out method. Various sets of stylometry and ad hoc social networking specific features were developed to categorize short messages from thirty Facebook authors as authentic or non-authentic using SVM.

The challenges of applying traditional stylometry on short messages were discussed. The full set of 233 features achieved the best accuracy rate of 79.6% over any of its subsets. The social network-specific features showed marginal accuracy improvement when added to stylometric features. However, users who adopted these features were more distinguishable in writing styles. The test results showed the impact of sample size, features, and user writing style on the effectiveness of authorship authentication, indicating varying degrees of success compared to previous studies in authorship authentication in short text. The proposed stylometric features and method were also tested on 300 sample long book data. SVM showed better accuracy rate than k-Nearest Neighbor (k-NN) on Facebook data, while k-NN showed a better accuracy rate than SVM on book data. Finally, a comparison of a number of commonly used classification methods was tested on Facebook data to assess their performance for short text authorship authentication. Decision tree showed the best accuracy rate followed by SVM with a linear kernel function.

Acknowledgements

I would like to express my gratitude to Dr. Li-Chiou Chen, in particular, for her guidance, encouragement and patience. I would not have achieved this milestone without her guiding me in the right direction. She gave me new hope and insight whenever I felt lost in this long journey. I felt hopeful every time I talked to her, which gave me energy and confidence to keep moving forward. With every small step I took, I was getting nearer to my goal. I could not ask for a better advisor.

To Dr. Charles Tappert, thank you for your advice and collaboration in this study. Your guidance was essential.

To Vinnie Monaco, I sincerely thank you for all your contribution and effort in the long text comparison. Your contribution added important value to this research.

To Dr. Fred Grossman, thank you for your guidance in each stage of the doctorate program.

To Pranjal Singh, thank you for your contribution of this research. Your results brought interesting insights.

To the students who helped me with the data collection, thank you all for your help in making this research possible.

To Chris Longo, thank you for taking care of all the administration questions I came across. You are so helpful.

To my classmates at Pace University, you are my support system. Thank you for marching towards the finish line with me. A special shout out to Sunil Rajan and Jean Chu. Thank you for checking on my vital signs periodically.

To my mom, sister, family and friends, you are my energy source. Thank you for believing in me. Your encouragement energized me throughout this long journey. You picked me up when I was close to losing faith and reminded me that giving up was not an option.

To IBM, thank you for your sponsorship.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Social networking sites, such as Facebook, MySpace, Twitter or Instagram attract millions of users. These sites provide environments for users to connect with their friends and family, or even to make new friends. Social networks resemble a virtual communication medium or are like online communities [68]. Users get together in these communities for information sharing or to build relationships. While users may assume social networks provide a trusted environment for sharing information with friends and families, the information maintained by the social network sites can be compromised. In 2010, 0.06% of 1 billion (600,000) Facebook logons were compromised daily [92]. For example, hackers could spam users with harmful messages, or hack into users' accounts and post fake messages on the users' behalves [13]. These malicious posts often lure users to click on a URL that leads to an external website. The URL link could include malicious code that can potentially steal users' information from their computers or mobile devices [29]. Others who may not suspect the authenticity of such fake messages may respond to them by clicking on the URL or respond with their opinion, preferences or even personal information. This could also be harmful, as intruders with malicious intent can now gain insights of others' information, which can potentially lead to financial loss or even identity theft, which is the top social network security threat [29]. Authorship authentication was one of the top trending security concerns in social

networks ([29] and [84]). Social networking sites should be enhanced with mechanisms to differentiate authentic messages from the ones that are not, so that users are safe to share information and interact with each other.

1.1 Purpose of the Research

At the current state of the art, no prominent social networking site has provided any built-in feature that re-authenticates users once they have logged in. For example, Facebook is currently the most popular social networking site [36] - they currently use textual login and password to authenticate users. Once a user has logged in, there is no re-authentication or detection of any abnormal behavior of the user. Unlike business transactions such as purchase orders that users may create occasionally, social network users can create as many postings as they wish on a daily basis. For example, Facebook users can change their status, share pictures, share links, create or join events, create notes, comment on others' postings, subscribe to commercial pages, etc. Some users are more active than others. Active users may initiate multiple postings and make multiple comments on others' postings per day. It was estimated that 293,000 status updates were made per minute on Facebook [90] in October 2014. With the among of postings per user per day, adding a security feature such as digital certificate to each posting is overkill and is not practical in such a causal and heavily used environment. It will discourage users from sharing information. This creates new security challenges in social networking sites on how to re-authenticate users by their behavior including the way they write.

Social networking site users should be able to trust the source of messages posted in a social networking site by their friends or people they know on a regular basis. They should not have to worry about the authenticity of every message that they come across. For example, if John sees a message from a friend who asks for vacation recommendations, John would not suspect that the message is not real or suspect that his friend's account is compromised. John may respond to him with relevant information on his vacation preferences or experiences. John would then be in danger of releasing his personal preferences to a hacker. A hacker can disguise himself as that user, post messages, comment on a user's circle of friends' postings, organize events, or perform any number of other actions on the user's behalf. In a more serious situation, a hacker can post messages to lure others to give out personal information or ask for monetary donations. It is important for social networking site service providers to provide some security mechanism to authenticate the authorship of messages or to detect any message that does not conform to the writing style posted by the same user. In addition, social networking site providers can provide awareness to the viewers if a message looks suspicious. Depending on the access control set by each user, a suspicious message from an individual's account can be viewed by friends only, friends of friends or even by everyone in the social networking site. A suspicious or harmful message can be potentially viewed from one circle of friends to another. This is similar to spreading germs or viruses.

The purpose of this research is to assess authorship authentication using short text. The research included messages collected from Facebook, which tend to be short in nature [46]. Facebook is currently the biggest social networking site [36], and is used herein to

illustrate and assess the solution proposed in this research. Leveraging users' writing styles, this research identifies means to differentiate authentic posts from unauthorized or fraudulent posts in social networking sites that are not authored by the real users who are associated with them. In short, the purpose of this research is to find a solution to these questions in social networking sites:

- How can we tell if a message is posted by the real user and not by others who hack into the user's account?
- Can I trust the authenticity of a message posted by my family, friends, or anyone in my social network?
- Is it safe for me to respond to a message in a social networking site?
- Can we develop a user's profile by the way each individual writes in a social networking site?
- Do users write differently in social networking sites?
- How accurate would it be if each user's style of writing in a social networking site is used as his or her biometric?

1.2 Scope

While social networking sites have gained tremendous popularity, security threats to users have also become more pervasive [69]. Spam, flaws in third-party applications, worms, and phishing are just a few of the sample attack methods that hackers can use to gain information from others, which could potentially lead to identity theft and other associated personal losses. This research addresses a fundamental trust concern of whether messages are posted by legitimate users - or not. It addresses authorship

authentication through investigation of users' writing styles. Given a set of available messages posted by a user, this study will determine if a new disputed message was authored by the same user. Various researches ([5], [16], [80], [91] and [108]) done on authorship identification, which differs from the goal of this research. Authorship identification is about assigning the correct author for a disputed message after investigation. This research, however, focuses on whether the message is authored by the user whose name is associated with the message when it is posted. It is an authorship authentication problem, with a binary "Yes" or "No" response for the question "Was this message actually posted by the user?"

For the scope of this research, we were interested in studying authorship authentication for social networking site messages, which tend to be short messages compared to books, articles, emails, or blogs. We used Facebook messages to demonstrate authentication with short text. Hussain et al. [46] studied Facebook data from three prominent Danish politicians' Facebook walls over a two-year period from October 1, 2009 to September 15, 2011. A Facebook wall is a user profile page where the user can post his/her own status, and others can leave messages there as well. A total of 162,646 posts, made by 25,987 individuals from these walls, were collected. The average number of words per post was 39.9. Hussain et al. have verified that Facebook messages were short in nature. In this research, we defined short messages as 50 words or less.

Each user has his or her style of writing. As social networking sites provide friendly and causal environments for people to connect with their families and friends, users have more freedom to express themselves in different style of writings. For example, some users start a message with the word "I", while others may speak in a third person's voice.

Some users like to use different punctuation marks while others may use common ones such as "," or ".". Users who speak multiple languages may even include different languages in the same message. Some users may use abbreviations like "LOL" ("Laugh Out Loud"). Was the message written with proper grammar? Does it contain a lot of punctuation or emoticons? Does it contain any foreign words? All of these are indicators that show how a message can be presented in many ways depending on each user's writing style. Through the study of each user's historical postings, individuals' writing styles can be identified and analyzed.

This research provides a mechanism to detect whether messages shared on a social networking site were authored by legitimate users and not by hackers. Even though most social networking sites allow users to customize their access control to people who have been granted permission to access their content, these restrictions are set on the viewers' side. A security gap remains from the content providers' side, as there is no existing built-in technology available on social networking sites to determine the authenticity of messages posted. This study fills in the gap, helping viewers verify that messages that are presented to them are authentic and safe for them to leave comments.

Social Writing Style (SWS) is the proposed solution in this research, using the combination of stylometric and social network specific features to assess the writing style of authors with their short text, Support Vector Machine (SVM) with a linear kernel function as the machine learning tool, together with the Leave-One-Out cross-validation method to detect the authenticity of messages in a social networking site. This research investigated whether users adopted a more causal writing style in social networking messages. The solution proposed can be used as a secondary authentication mechanism

by detecting abnormal writing behavior that deviates from individuals' usual writing style. 9,259 Facebook posts were collected from thirty (30) users and were tested to demonstrate the idea of SWS for authenticating users with short text. The Leave-One-Out method was used to overcome the shortage of samples from a number of users. The accuracy rate of using SVM as the machine learning and classifier method was provided. The proposed features and method were also tested on book data to assess the accuracy rate when used for long text authorship authentication. Finally, a number of popular classifiers were tested on Facebook data to assess each accuracy rate for authorship authentication in short text.

1.3 Research Methodology

Facebook was used as an illustration for this study, as it is the most popular social networking site [114] and it served as a great data source for online messages that tend to be shorter than articles or documents [46]. This study followed a traditional research methodology on classification. Thirty (30) Facebook users' status posts for a period of four years were collected manually by copying and pasting each user's messages into a text file with each message starting on a new line. The decision to collect data manually was made after considerations of the substantial effort required to learn Facebook application development if data was to be collected automatically. The list of users consisted of six friends who agreed to participate in this research, and 24 public figures whose Facebook profiles and postings are available for public viewing. The research methodology involved the following steps:

1. Design of features
2. Data/sample collection

3. Features extraction from samples for each user
4. Composition of training and testing data for each user
5. Training of sample training data for each user with machine learning software
6. Classification of testing data for each user with machine learning software
7. Validation of classification results by assessing the False Acceptance Rate (FAR), False Rejection Rate (FRR), Equal Error Rate (EER) and accuracy rate

To assess the writing style of each user, this research leveraged stylometry to assess the authenticity of messages. Stylometry is the study of linguistic style, usually to written language [116], and is measured in stylometric features [49]. In this research, a selected list of 233 features that included 227 stylometric and 6 social network specific features were designed as measures for each post. The list of social network specific features was used to capture specific causal writing styles for online messages including the use of emoticons, internet slangs, and other identifiable attributes. They were designed to assess the accuracy of classifying authentic messages from others when the authors adopted a more causal writing style.

An AWK program was developed for feature extraction. It read the individual input files that stored sample Facebook posts for each user, calculated the feature measurement for each post repeatedly for all posts collected per user, and generated a corresponding vector file (text file) representing those feature measurements. Each message was measured against the 223 features, and each feature and its associated measurement were represented in a name value pair. Each line of the vector file contained 233 name value pairs, which showed each of the 233 features that were measured. Each line of the vector

file represented the measurement of a message, and each vector file represented the feature measurements of messages collected from a user.

After the feature extraction was completed, training data and testing data were formed for each user. For each user, the training data would be used by a machine learning software for feature analysis and development of a model that represented the target user's writing style. In order for the machine learning software to accomplish this, both samples from the target user, as well as samples from others, were needed during training so that it could analyze which writing characteristics defined the target user and, perhaps more significantly, which ones did not.

Support Vector Machine (SVM) was selected as the machine learning and classification method as it showed a great accuracy when compared with other classification methods from researches ([80], [91] and [122]) done in authorship studies. SVM Light, an implementation of SVM [54], was selected as the machine learning and classification software as it was both well-documented and publicly available. SVM Light provided a "training module" to train each user's training data and develop a user's model (a profile that represented the user's writing style) based on feature measurements of that training data. SVM Light provided a "classification module" to assess the testing data of the target user against the user's model to determine if each message in the testing data set was authored by the target user or not.

The output of the classification was a prediction of whether each testing data/message was authentic or not. A positive number indicated the testing data was authentic, while a negative number showed otherwise. After all users' testing data was classified,

performance in term of False Acceptance Rate (FAR), False Rejection Rate (FRR), Equal Error Rate (EER) and accuracy rate were calculated. FAR represented the rate of accepting unauthentic messages; FRR represented the rate of rejecting authentic messages. Both are errors that showed data/messages were classified incorrectly. The EER was derived from the average of FAR and FRR. The accuracy rate was defined as 1-EER. Additional details on the research methodology will be discussed in Chapter 5.

1.4 Audience

This research focuses on answering the question "Is this post authored by the real user?" Can the readers trust the message to be authentic? The research provides a means to predict if a message is written by the real user, based on the writing style developed by the same user from his/her previous postings over a period of time. This research could provide a way to re-authenticate a user when new messages are posted. Researchers who are interested in improving the security of social networking sites should be interested in the findings of this research, and they can extend the proposed solution to improve the accuracy rate. In addition, there are two types of audiences that may also benefit from this: social network service providers, and social network users. Social networking site providers have the responsibility to ensure a safe and friendly environment for their users as they share their information as frequently as they would like on a regular or daily basis. Social networking site providers can use the proposed solution as a security mechanism to automatically detect potentially malicious hacking by proactively monitoring the authenticity of messages, protecting users from exposure to messages that may be fraudulent or harmful. When a suspicious post does not conform to the user's writing style, the social networking site provider can verify the authenticity of the

message with the user, or provide alerts to the user's circle of friends to warn them ignore the message. Also, the proposed solution can serve as an additional security mechanism to protect each user's account from malicious attack, since users whose accounts are compromised, along with their circle of friends, can be notified to reduce potential harm.

1.5 Summary

Once a user is authenticated with a social network, there is no re-authentication. If a hacker is able to log into a legitimate user's account, he or she can perform a myriad of activities including status updates, sharing links, pictures or video, creating events, etc. Others may not notice these activities were not performed by the real user. If a circle of friends respond to fraudulent postings, they are in danger of providing personal information to the hacker, propagating the damage, since the hacker can potentially steal the identity of people who he/she now has established connections with. This study attempted to address the security gap of current social networking sites by providing a potential secondary authentication mechanism, verifying authenticity of messages posted based on individuals' writing style.

Each user writes uniquely. Users may consider social networking sites as friendly and casual environments for them to share information. Some may adopt a more casual writing style, which is different from the more formal writing style utilized in letters or publication. Data from thirty Facebook users was collected in this research and used as a demonstration of how users write in a social network. Specific stylometric and social network features were selected and extracted from each user's samples to develop a

profile or some sort of biometric that can identify the user from others. Support Vector Machine (SVM), a popular machine learning and classification method, was used to evaluate the performance of the selected features. The Leave-One-Out method was used to overcome the issue of insufficient samples from a number of users. In addition to detecting fraudulent messages that were not authored by the real user, this research provides a comparison of performance of authorship authentication with short messages verses long messages in term of accuracy rate, and shows which machine learning classifier yielded a better accuracy rate with short messages. Results of this study are discussed in Chapters 6, 7 and 8.