Application to Identify the True Sender in Instant Messaging

Hanaa Mohsin Ahmed and Shahad Fadhil

Computer Science - University of Technology – Baghdad - Iraq

Shahad_fadhil94@yahoo.com

Received: 5 November 2017 Accepted: 3 January 2018

Abstract

As cybercrime continues to spread, new e-forensic technologies are needed to counteract persistent yet anonymous fraud. The anonymity offered by the Internet has made the task of tracing criminal identity more difficult. Criminals use virtual identities to hide their true identities, for example via instant messaging (IM), which conceals inherent security risks and malicious tendencies from standard security software. It is essential that there are electronic forensic techniques to help identify cybercriminals as part of a criminal investigation, including identifying gender, writing habits, and stylometrics. This research paper uses statistical methods for the analysis of a template matching Pearson correlation coefficient or Euclid similarity to analyze Enron data and Viber chat messages with regard to distinct data in terms of the definition of the sender using neural networks. We also describe the structure of the creation and analysis of stylometric features for the cybercrime for electronic crime investigations. The proposed system has found a high rate of true identification of the real IM sender, using Pearson matching (88%) and an acceptable rate for Euclid similarity (72%).

Keywords: Instant Message, Identification, Pearson correlation coefficient, Euclid similarity.

Application to Identify the True Sender in Instant Messaging

Hanaa Mohsin Ahmed and Shahad Fadhil

تطبيق لتحديد المرسل الحقيقي من خلال رسائل الفورية

هناء محسن احمد الوكيل و شهد فاضل عباس

قسم علوم الحاسوب - الجامعه التكنولوجية - بغداد - العراق

ألخلاصة

الجرائم الالكترونية اصبحت تشكل خطر على المجتمع الالكتروني و وسائل الاتصال المتداولة في هذا المجتمع واصبحت في تطور سريع ومن الصعوبة تحديد المجرم الألكتروني. اصبح يتخفى وينتحل الشخصيات ويقوم بالجريمة من خلال الرسائل الفورية (رسائل التواصل المباشرة مثل الفايبر) وهي تمثل الخطورة المحتملة في الاختراق والتجسس. تستخدم التحليلات الجنائية وسائل وطرق كثيرة لتحديد هوية المجرم في مسرح الجريمة لغرض التعرف عليه. ومن هذه الوسائل تحديد جنس او التعرف على اسلوب كتابة وصيغة كلامه. وفي هذا البحث تم استخدام طرق احصائية وطرق الاختبار ومن ضمنها الطرق الذكية في الشبكات العصبية, وكذلك بناء مخطط لعملية تحليل للنص والتعرف على اسلوب الرسالة المرسلة وحفظ المعلومات حول تلك الرسالة. تم استخدام بيانات اينرون العالمية وكذلك الفايبر لتجريب وفحص هذا النظام وكانت نتائج أداء النظام في استخدام اقليدس هي (72%) وفي طريقة مطابقة بيرسون هي (88%) .

الكلمات المفتاحية : الرسائل الفورية، تحديد الشخص، تحديد الاختيار الصحيح، تشابه اقليدس.

Introduction

The rapid development of the Internet has created countless ways to exchange information, online social networking, interpersonal communication, and the promotion of goods/services, events (such as Viber, WhatsApp, Twitter, Facebook, and so on), as well as e-commerce (such as Amazon and Craigslist) [1]. This growth has contributed to the enhancement and enjoyment of life. However, it also has been abused. The anonymous nature of the Internet allows online criminals to use virtual identities to hide their true identities to facilitate their crimes [2]. Many users have encountered the problem of anonymous documents and texts counterfeit authorship. As the utilization of web advances are expanding, the use of instant messaging (IM) is also expanding, and investigating these mass messages is a difficult undertaking for forensic analysts [3].

The goal of this research paper is to analyze stylometric IM to identify the sender in cases of impersonation and the violation of senders' rights.

Related work

In the literature, many researchers have presented works on either person identification or gender identification. Below are some of the last years of researchers' work and their use related to this research paper:

1- In 2014 [1], Na Cheng, R. Chandramouli, and K. P. Subbalakshmi presented a paper that focused on identifying the gender of the author of a text because some people fake their gender in Internet Correspondence. They used a Support Vector Machine (SVM) classifier and a Bayesian logistic regression classifier, which depended on linguistic and writing styles, as well as certain words that are commonly used. They proved that the SVM classifier appears to be the best candidate for their gender identity in the text.

2- In 2015 [2], V. Sreenivasulu and R. Satya Prasad proposed semantic ontologies Gaussian mixture model of data mining models to investigate cybercrimes to analyze bulk emails gathered forensically, depending on the semantic tool to analyze the grammar, and using metrics such as the False Acceptance Rate and the False Rejection.

3- In 2016 [3], Mercy D'cruz A suggested a maximum likelihood algorithm in the context of short texts that identified the maximum likelihood of a particular author has written a Section of text by examining other sample texts produced by that author. The extraction features that were used for experiments were character related, digits, punctuation, special characters, word related, functional words, and parts of speech.

1. In 2017 [4], M. Al-Smadi, Z. Jaradat, M. Al-Ayyoub, and Y. Jararweh paraphrased identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features. This research proposed a state-of-the-art approach to paraphrase identification and semantic text similarity analysis and used Support Vector Regression With extracted features, which were lexical, syntactic, and semantic. The results showed that the approach achieved good results in comparison to the baseline results by 0.841 paraphrase identifications and 0.892 semantic text similarities. And bellow

collected the conclusion about previous work The (lexical, Emoticons, Syntactic, and Stemming) features were usually a suitable to be used.Identification methods could be of gender or author style, depending on Enron chatting messages database and used Viber message as implement the proposed system. Extraction features came in two stages, the first stage that was identifying gender. The second phase extracted the rest of the features, to determine whether the person was himself or impersonate.

The Proposed System

The proposed system is used to identify the sender using stylometric concepts by identifying the gender and the writing style of the person. The proposed system is based on using chat messages as letters written in the English language; see Figure 1. The system consists of two stages that form the basis of the system's operation. These stages are:

Stage 1: Gender (detection by using a look-up table).

Stage 2: Stylometric (calculates features to determine the writing style of the sender), and work requirements to produce accurate results.

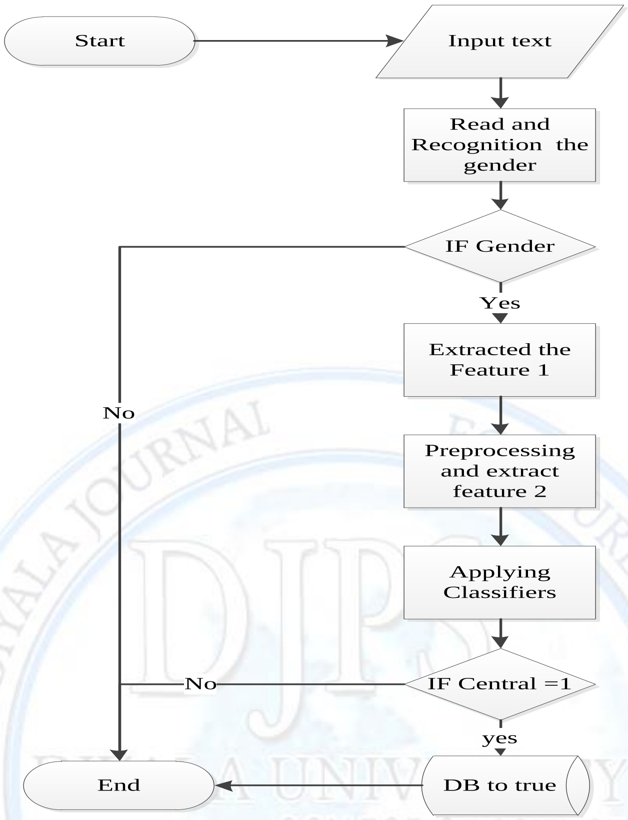

Figure 1: Flowchart of the Proposed System.

1- Input Message

A data set containing chat messages was needed for a group of people for training purposes. The set was taken from Enron chat messages [5], which made it easier for us to extract writing patterns for each person. In addition, we took chat messages from the Viber application as a test by extracting the backup of a message.

2- Gender Identification

This feature was created to determine whether a person is male or female. This is considered one of the most important features. It must be able to pass all words in the matrix cell and compare them to a set of stored male or female words; these words were stored using a table as a dictionary database.

3- Stylometric

This is the basic and critical step in the stylistic analysis of chat messages, in which they must be configured as a suitable input format. This step is crucial in determining the quality of the next several steps; the stylometric phase involving tokenization is involved, representing the step of the stylometric phase, followed by the phase of selecting features, and the classification of the steps.

4- System Analysis

4.1 Data Preprocessing

An essential stage in initiation attribution. Content records in their unique shape are not in the appropriate frame for learning. They should be changed over to a reasonable information organize. It can be changed over into a vector space since the vast majority of the learning calculations utilize the characteristic esteem portrayal. This progression is significant in deciding the nature of the following stages, that is, the component extraction and order organize. Information pre-preparing includes tokenization and stemming. 1.1

A-Tokenization

This is the process of segmenting the text into a collection of features of meaningful elements that are called tokens, such as symbols, phrases, words. The tokens are extracted about as a contribution for further procedures such as parsing and data mining. Each text is divided into sentences based on the “.”, each sentence is divided into various tokens based on the spaces between words.

B- Stemming

Stemming is a method for identifying the root form of the word and forms of derived words. Because derived words are semantically similar to their root forms, this retains the meanings of words to improve classification by removing additions or affixes (prefixes and suffixes) from words.The aim of stemming is to reduce all variants of a word to a single term to extract similar words, and to improve the performance and efficiency of the system when the various forms of tokens are stemmed into a single form. The stemming methods lead to dictionary words to detect the words that must be removed or edited according to the rules of grammar; thus, it is necessary to produce a dictionary database, which is done by Algorithm (1), Process the sentence and return the words to their root with tokenization and stemming:

```
Algorithm (1) : Processing of the sentence
Input     Sentence as array of word
Output    Sentences with token and stem
Begin
  Step 1: Delete numbers, special characters and symbols.
  Step 2: Remove stop words.
  Step 3: Convert all large characters to small characters.
  Step 4: Reduce all variants of a word to a single term or root; for example, {drink, drank, drunk}
          → “drink” It is generally sufficient that related words are linked to a similar stem and
          representative samples of irregular verbs, irregular adjectives, irregular nouns, and
          irregular adverbs.
  Step 5: Additions or suffixes (prefixes and suffixes) are removed from words by a set of simple
          rules to find stems. These results are considered features through which the identity of the
          author is determined.
End
```

4.2 Feature Extraction

After performing the feature extraction process, the extracted features are used to classify the input text data and Enron data. Mainly two classifiers are used and they are Backpropagation classifier and Pearson template matching
classifier [6].

A- Features Before Stemming

The features that were selected before text enter in preprocessing that means represent an aspect of a personality stylometric feature for the user. Features and extraction processes depend on the text language. These specifications can be used to understand the writing style of the sender, and are extracted from the sender’s messages. And compared with the feature that extraction after Stemming preprocess. Numerous types of features have been used in previous studies to analyze authorship attribution based on the author’s writing style, including lexical, syntactic, structural, and functional word-based features. A string of characters and symbols is formed and then we extract the features. The following important features are extracted:

1. Number of Letters (N.L.)
2. Number of capitals (N.C.)
3. Number of specials (N.S.)
4. Number of numbers (N.N.)
5. Number of smiley faces (N.F.)
6. Number of abbreviations (N.A.)
7. Number of words that are similar (S1).

B- Features after stemming

In this steps of the features are extracted, the repetition of existing words, which have been identified after the process of stem, which in turn restoration words to their root. The calculation of the repetition of words and compare with the repetition of words in the first case and take samples for each person distinguished from others by style writing. Called this feature by similar 2 (S2) for represented results of the step.

4.3 Applying Classifiers

A- Euclidian Similarity Matching

The Euclidian distance (Ed) is the basis of many measures of similarity and dissimilarity. It examines the square root of the sum of squared differences between corresponding elements of the two vectors. The element of each vector can be standardized variables, such as [7].

Step 1: calculate the mean value of each column

$$mean = \sum_{i=1}^{n} \frac{x_i}{n} \qquad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Step 2: Calculate the standard deviation

$$standard\ deviation\ (std) = \sqrt{\sum_{i=1}^{n} \frac{((x_i - meanx_i)^2)}{n}} \qquad \ldots\ldots\ldots (2)$$

Step 3: Calculate the normalization (normalize the column vector by subtracting the mean value from each vector and dividing by the standard deviation)

$$N = (x - meanx_i)/std \qquad \ldots\ldots\ldots\ldots\ldots (3)$$

Step 4: Calculate the mean value for the new columns (N)

$C = mean(\left(xi - \frac{meanx}{Std}\right))$ …………….. (4)

Where $x = orginal\ value.$

The figure (2) shows the central value of user that calculated the central value.

The Euclidian distance equation is

D (q , p) =$\sqrt[2]{\sum_{i=1}^{n}(qi - pi)^2}$ ……….(5)

B- Classification using Back Propagation NN

Backpropagation auto-associative is a feed-forward neural network that was trained but is not the same as a typical neural network. The conjectural training phase teaches the network to output vectors close to the input training vectors. During the test phase, the input vectors that result in a difference from the outputs are designated as having a high degree of anomaly. In the training phase, the neural network is constructed with feature input nodes and feature output nodes. We used [2 × feature/3] = 4 hidden. The network is trained to supersede each input training vector as the output. We trained for 300 epochs using a learning rate of 0.0001 and a momentum parameter of 0.0005. In the test phase, the feature test vector is run through the network, producing a feature output vector.

C- Pearson template matching

The Pearson correlation coefficient was chosen to measure the direction and magnitude of the relationship between two variables X, Y. The quantitative measure of the degree is the correlation coefficient, which always ranges between –1 and +1 [8].

$A\ and\ B$. Eq (6)

$$Asub\ and\ Bsub = x - mean_{A\&B} \quad \text{……. (6)}$$

$$covariance\ AandB(covAB) = mean(Asub \times Bsub) \quad \text{…….. (7)}$$

Then find:

$stdA$, and $stdB$ according to Eq (2)

Following this, we can find the correlation coefficient ρ

$$\rho_{A,B} = \frac{covAB}{stdA \times stdB} \quad \text{…… (8)}$$

Results and comparison to previous work

Relative frequencies for all style markers for each text (chat message) were generated by means of a tool programmed in Visual Studio C#. The chat messages extracted from the Enron dataset and the Viber messages were analyzed, and each message was processed using concepts such as tokenization and stemming, with each of the words having different synonyms, the figure (2) shows the separated central value for every user, it was easy to detected the user by its line in range value. Each line represented the feature extraction from the message and calculated by standardized variables equation (1,2,3,4). A field was used to extract the meaning of each of the words in which gender was determined. For the evaluation, we selected messages from an Enron dataset composed of 15 users 20 chat messages for each. Most of the email messages were from males and some were from females. For this evaluation, we selected for gender first, and seven features were extracted to detect to whom the chat message belonged. In this paper Selected various random chat messages and showed what percentage of a chat message belonged to a central value by back-propagation methods.

Table 1 shows the results obtained from the back-propagation classification techniques. The first column detected the female/male. The next seven columns are the features extracted to provide more details about the dataset chat message. For the evaluation, we calculated a stylometric dissimilarity among all central dimensions to choose the highest percentage that matched the central classes. A thorough evaluation of a detected chat message would require the most important evaluation, as in the following:

- Calculate the number of features
- Calculate the central value and distribution for the user
- Training and testing back-propagation algorithms by central values
- Test by template matching
- Compare results of the detected message for identification.

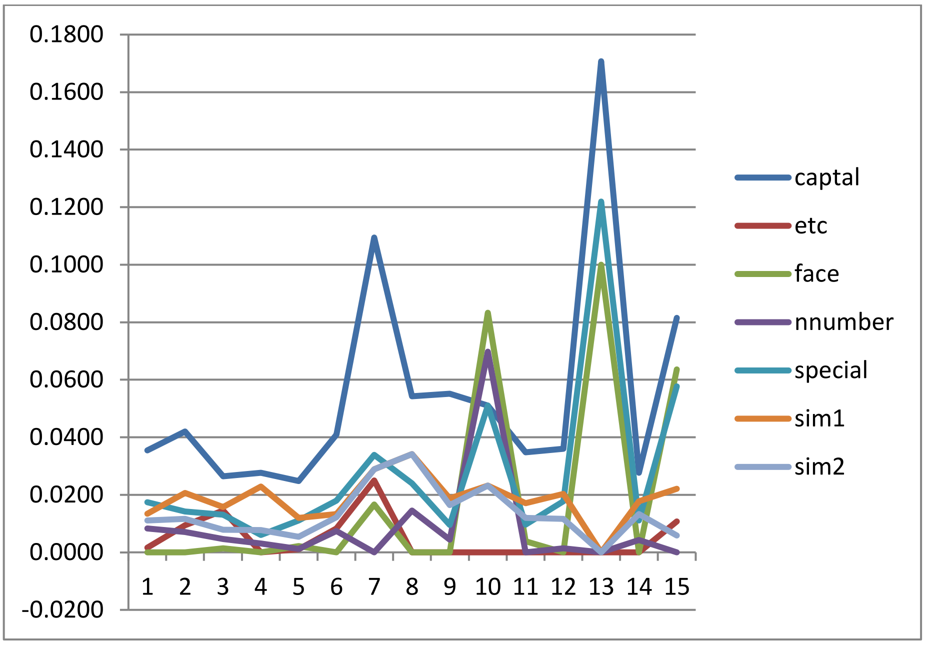

Figure 2. Separated central value dimension

Table 1: Classification of chat message according to feature

message	G	N.L	N.C.	N.F.	N.S.	N.N.	N.A.	S1	S2	Recognition
Congratulations, :) Will! May God :) bless you and your family She's a beautiful baby. How is recruiting going? Any idea where you think you will end up. I'm sorry it did not work out at Enron. Know that you will do well wherever you go.	F	178	10	2	7	0	0	5	2	1 = 83% 3 = 40%
If you could start thinking about these in greater detail, we can then go through and incorporate into the west orig. objectives. Call me to discuss, but we should start to discuss detail early next week.	F	165	2	0	2	0	0	2	2	1=30% 2=75%
Don't get too excited to hear about the rest of the day...I don't have much positive to write. It's been boring and I'm very uneasy about my role in this group. More later…	N/A	128	5	0	0	0	0	2	2	1=25% 12=81%

Table 2: show Summarize the results obtained after the trials. the Percentage value of classifiers

S.NO	Classifier Type	Percentage value
1	Euclid with BP	72
2	Pearson template	88

Conclusions and future work

In this research paper, the sender of the IM was identified by first using gender identification, followed by stylometric identification. The program was created based on the classification and selection of the sender through the writing Pattern. Experimental results showed that the extraction of the appropriate traits through the consideration of linguistics and writing style was adequate and of sufficiently high importance to Identify the person. The template matching for Euclid similarity was 72%, but the Pearson correlation coefficient (88%) was more effective and more accurate for distinguishing between persons. And therefore,

1. After applying the suggested method, we found that, in the case of using Viber, the ratio of excellence was 84%; in the case of using Twitter the percentage of excellence was 80% and in Enron data, excellence was 88%.
2. In this paper, the results compared the practical section with the above for previous works. The number of features used has been reduced from 356 to 8.
3. Using the proposal in this paper, the results of excellence in detecting males and females have been shown to be 85–95% accuracy in detecting females.
4. For future work, can use biometric features to prove identity more precisely by integrating several methods

Reference

1. A. M. Rezaei, “Author Gender Identification from Text,” Eastern Mediterranean University, 2014.
2. P. Sreenivasulu, R. Satya Prasad, “A Methodology for Cyber Crime Identification using Email Corpus based on Gaussian Mixture Model,” Int. J. Comput. Appl., vol. 117, no. 13, pp. 29–32, 2015.
3. B. Sivakumar and K. Srilatha, “Maximum Likelihood Text Classification Algorithm Using ML for Authorship Attribution,” Int. J. Innov. Res. Comput. Commun. Eng., vol. 7, no. 3, pp. 365–373, 2016.
4. M. AL-Smadi, Z. Jaradat, M. AL-Ayyoub, and Y. Jararweh, “Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features,” Inf. Process. Manag., vol. 53, no. 3, pp. 640–652, 2017.
5. S. NagaPrasad, V. B. Narsimha, P. Vijayapal Reddy, and A. Vinaya Babu, “Influence of lexical, syntactic and structural features and their combination on Authorship Attribution for Telugu Tex,” Int. Conf. Intell. Comput. Commun. Converg., vol. 48, no. C, pp. 58–64, 2015.
6. E. E. Abdallah, A. E. Abdallah, M. Bsoul, A. F. Otoom, and E. Al Daoud, “Simplified features for email authorship identification,” Int. J. Secur. Networks, vol. 8, no. 2, p. 72, 2013.
7. Lakshmi and P. K. Pateriya, “A Study on Author Identification through Stylometry,” Int. J. Comput. Sci. Commun. Networks, vol. 2, no. 6, pp. 653–657, 2012.
8. S. Barate, C. Kamthe, S. Phadtare, R. Jagtap, and M. R. M. Veeramanickam, “Text Character Extraction Implementation from Captured Handwritten Image to Text Conversionusing Template Matching Technique,” ICAET, vol. 1010, 2016.