Application of Stylometry to DarkWeb Forum User Identification

Thanh Nghia Ho(✉) and Wee Keong Ng

Nanyang Technological University, Singapore, Singapore
hoth0002@e.ntu.edu.sg, wkn@pmail.ntu.edu.sg

Abstract. The fast growth of the cyberspace in recent years has served as a convenient channel for criminals to do their illegal businesses, especially in Dark Web - the hidden side of the Internet. The anonymous nature of Dark Web forums makes them ideal environments for criminal discussions. Ranging from government, security agencies to financial institutions, many parties are willing to trace the identities of the suspects through these online conversations. Dark Web participants usually have multiple accounts on various forums. On multiple occasions, being able to validate that multiple accounts on different Dark Web forums belong to the same person with high enough confidence allows us to combine various scattering pieces of information into a more concrete and advanced form of knowledge. Such knowledge will lead to actionable insights which are very useful for bringing the criminals to justice. In this paper, we examine the effectiveness of writing style analysis (stylometry) for linking multiple accounts in different Dark Web forums. Initial evaluations have shown that the proposed methodology is promisingly practicable, having a high potential to assist the investigators in exposing anonymous identities in cyber environments.

Keywords: Dark Web · Stylometry · Support vector machine

1 Introduction

Cybercrime is the act of exploiting Internet resources to commit illegal activities ranging from identity theft, drug sale, credit card stealth, and child pornography... It has been shown through past experiences that dealing with this crime wave is not a trivial task. The anonymous nature of the Internet and the ease to enter online criminal markets have greatly contributed to the dramatic increase in cybercrimes. According to Singapore Police Force (SPF) in its annual crime report in 2015, online commercial crimes are the main factor in the 4% rise of overall crime in Singapore. The existence of Dark Web further enhances the anonymity of online criminals. Unlike Surface Web, Dark Web can only be accessed using the TOR browser which provides a mechanism to hide surfing users' information from the Internet Service Providers (ISP). As a result, law enforcement agencies have difficulty in gathering as much information about the criminals on the Dark Web as in the Surface Web. However, there is no such

K.-Y. Lam et al. (Eds.): ICICS 2016, LNCS 9977, pp. 173–183, 2016.
DOI: 10.1007/978-3-319-50011-9_14

thing as a perfect crime. The criminals may unconsciously leave some trails on the Internet in various forms of information, e.g. texts, images, and videos... These trails can provide valuable information which may help to reveal the criminal identities. In this research, we mainly focus on the analysis of criminal textual data in Dark Web forums. Specifically, we aim to profile Dark Web users by linking up their accounts on multiple Dark Web forums through writing style analysis. We consider, for example, a criminal selling drugs publicly in one Dark Web forum who also has another account with partial personal information on another Dark Web forum. The police can take advantage of that combined knowledge to carry out legal actions on the criminal.

In Sect. 2, we review the previous works that tackle the problem of authorship attribution. Section 3 provides an overview of the dataset that we use in this research. In Sect. 4, we formally introduce the aforementioned problem of authorship matching between users on multiple Dark Web forums. We present the main work, our novel authorship attribution algorithm, in Sect. 5. In Sect. 6, we evaluate our algorithm on a dataset collected from multiple Dark Web forums. Finally, in Sect. 7, we give a conclusion for this research and discuss a few potential future works.

2 Related Work

2.1 Dark Web Forums

Traditional Web search engines, including the powerful ones, such as Google and Bing, are unable to index everything on the Internet. In general, we can divide the Internet into two main parts: the visible layer (Surface Web) and the invisible layer (Deep Web). The Surface Web contains all searchable contents that are visible to the general public. In contrast to the Surface Web, the Deep Web is the section of the Internet whose contents have not been indexed by standard search engines. The Dark Web is classified as a small portion of the Deep Web that has been intentionally hidden and is inaccessible through standard web browsers such as Google Chrome, Mozilla Firefox, and Internet Explorer...

Dark Web forums are utilized by cybercriminals to set up secret trading networks and support the exchange of illegal goods and services. Criminals use Dark Web forums' screen names or aliases to communicate with others. In addition, the user's private identification information is usually not shown on those forums and there are no restrictions on the number of accounts a single user can create. In other words, a user can create several anonymous accounts on the same forum and this kind of feature creates multiple challenges to the security and intelligence organizations in tracing back the identity of the anonymous users. For example, on 3 November 2014, the then 18 years old Newcastle man Liam Lyburd was arrested because of his threat to carry out a mass-killing at Newcastle College on a Dark Web forum called "Evolution". The police discovered that he had used two different accounts with unique usernames ("The Joker" and "I Love My Anger") to express his admiration for Anders Breivik, who was sentenced 21 years in prison for killing 77 people in Norway, and Jaylen

Fryberg, who carried out a college attack in the US in which four students were killed [1].

The Dark Web and its danger to the society have extensively raised public awareness in recent years. A great number of parties, including the government and academic researchers, have paid serious attentions to this online threat. Although it is difficult to detect the activities of Dark Web participants, the hidden information in Dark Web forums represents a significant source of knowledge for law enforcement agencies. Accordingly, many researches have been carried out to study the Dark Web forum data. For example, Abbasi et al. [2] introduced a sentiment analysis framework to classify Dark Web forum opinions in both English and Arabic. Zhang et al. [3] created a Dark Web Forums Portal to collect and analyze data from multiple international Jihadist forums. However, few attempts have been made on Dark Web forum user's data integration and association between different forums.

2.2 Attribution Techniques

Amongst the authorship attribution techniques, statistical and machine learning methods are the two most frequently used ones. The former has gained a good reputation for its high level of accuracy [4]. However, there are some complications in these approaches, including the need for more stringent models and assumptions. The exponential growth in computer power over the past several years has promoted machine learning as a better candidate to solve the problem. Machine learning consists of a variety of different techniques such as Bayesian, decision tree, neural networks, k-nearest neighbors (k-NN), and support vector machines (SVMs)... Wang et al. [5] pointed out two huge advantages of machine learning compared to statistical methods. Firstly, machine learning techniques are more scalable as they can handle a larger number of features. Secondly, an acceptable tolerant degree to noise can be achieved with machine learning techniques. These advantages are crucial for working with online messages, which often involve classification of many authors and a large feature set.

SVM classification is the most popular machine learning approach in recent years due to its classification power and robustness. Diederich et al. first introduced SVM to the online text classification domain in 2000 [6] where experiments are based on newspaper articles. After that, in 2006, Zheng et. al. [7] has shown that SVM is far superior compared to other classification methods for authorship identification of online messages.

3 Corpora

There are many challenges associated with the data collection on Dark Web forums compared to Surface Web forums. Accessibility is a big issue as most Dark Web forums restrict their contents to the registered users. They are known as *password-protected* Web resources that require registration and login. In addition, some Dark Web forums provide the contents in the unlinked form.

“Unlinked content” referred to the online pages which are not linked to by other pages. This helps to limit the capabilities of web crawling programs. Hence, creating an effective crawling tool for Dark Web forums is a time-consuming and uneasy task. For the purpose of this research, we use an existing data source that is available online. The bulk of our data was obtained from a vast archive compiled by Gwern Branwen, who is an independent Dark Net Market (DNM) researcher. The archive contains data from more than 30 different Dark Web forums which are mostly in the HTML format.

4 Problem Specification

We define the authorship matching problem as the task to validate whether two accounts having the same username on multiple Dark Web forums belong to the same person or not through writing style analysis. Perito et al. [8] discovered an interesting fact that users have the tendency to use similar usernames for their accounts in different sites, e.g., “donald.trump” and “d.trump”. Therefore, it seems that tracking different accounts of a user can be achieved by searching for similar usernames. However, this is not always the case. The chance that two different users coincidentally create accounts with the same username on two different forums is not rare. For example, two users who are the fans of Marvel comics/movies may create the accounts with the same username “spiderman”.

In this paper, we propose a novel approach for performing authorship matching using stylometry and basic classification technique (SVM). The problem can become very complex if the users purposely alter their writing style in order to make sure that they are completely anonymous. Kacmarcik et al. [9] have explored different techniques of automatically obfuscating a document to preserve anonymity and found that the number of changes per 1000 words is too small to confound the standard approaches to authorship attribution. For this research, we assume that the authors do not attempt to hide their writing style when participating in online discussions on Dark Web forums.

5 Experiments

Most authorship experiments usually follow a similar machine learning pattern: the model is first trained on a set of texts (training set) and then tested on another set of texts (testing set). The evaluation is based on the accuracy of the testing phase. We apply the described train-and-test process to examine the effectiveness of our proposed authorship matching techniques by designing two different author-matching tasks.

In the first experiment, we collect all the posts of a user in the same Dark Web forum and split them into two parts with nearly equal size and evaluate the authorship matching between those two parts using the proposed framework. This experiment is used as the validation process for finding the best parameter values. We label this process as “Validation Phase”.

After the “Validation Phase” is completed, we proceed to the “Testing Phase”. In this experiment, we identify some users that have accounts on multiple Dark Web forums. For each user, we collect his/her posts in two arbitrary forums and evaluate the authorship matching between those two parts.

5.1 Message Collection

In the first step, we need to define the list of authors that we want to analyze and extract a set of forum messages posted by those authors from the Dark Web archived data. As the archived data is in HTML format, a data transformation step is required to convert the HTML documents into raw plain text documents. In each post block, there are many redundant components that we need to remove such as user’s profile details, user’s signature, and quoted post of another user... The preprocessor first parses an HTML document and returns the actual user’s post contents. The parsing process is done with the help of “Beautiful Soup”, a python library for HTML parsing.

5.2 Message Filtering

Real-world classification problems are influenced by several components. Among them, the presence of noises is a key factor. Noise is an unavoidable problem, which affects the quality of the data collection and data preparation processes. The performance of the authorship matching model built under such circumstances does not only depend on the quality of the training data, but also on its robustness against the noisy data. The noise in text data is defined as the undesired blocks of words which provide no or very little information and need to be removed in order to enhance the total quality of the classification process.

For Dark Web online messages between users, one of the most common “noise” which can be easily observed is known as users’ PGP key. PGP stands for Pretty Good Privacy. PGP is most commonly used for data encryption and digital signatures. It adopts the public-key encryption mechanism in which a public key is used to encrypt the data and another separate private key is used to decrypt the encrypted data. Dark Web users usually make use of PGP to encrypt their sensitive information when doing illegal trading. Merchants on underground marketplaces usually provide own public key on their profile. Whenever other users want to trade goods or services from a merchant, they need to encrypt their shipping address using the public key provided by that merchant. In this way, only the merchant can decrypt and view the buyers’ messages.

Another common “noise” that we need to consider comes from the hyperlinks (urls) that appear in the user’s posts. Almost no user can memorize a long and difficult to remember url and retype it. In fact, users tend to copy/paste the url when they want to refer to an external source in their posts. Although each hyperlink is not directly typed by the authors, it still generates features for the stylometry process. For example, most of the feature set used in stylometry include bigrams and trigrams. The hyperlink usually begins with “http” or “https” and thus will create bigrams (“ht”, “tt”, “tp”) and trigrams (“htt”, “ttp”). These features contribute little to the stylometry process and may even affect the results badly if they dominate other features.

In addition, as our research mainly focuses on English text data, we need to remove the chunks of texts that are written in other languages. Language detection is performed using a Python module called “guess_language”, which adopts a trigram-based algorithm. “guess_language” can detect over 60 languages, including non-alphabetical ones such as Japanese, Chinese, and Korean...

Last but not least, we need to remove duplicated paragraphs from the text documents. For example, some vendors tend to spam advertisements about their products or some admins may post the forum rules and guidelines multiple times to remind the users. Another common duplication problem comes from the user’s forum signature. Most forums allow the users to add their unique signature in their user profile that is automatically appended to each of their posts. As a result, it reduces the exact length of the documents that are used in the classification. All the stylometry methods require some minimum number of words for both training and testing data. The more the duplications are present, the less effective the classification process become.

5.3 Feature Extraction

The purpose of the feature extraction process is to convert the raw text data into informative and non-redundant data representation to facilitate the subsequent learning step. The raw collected messages are in unstructured fragments, which are not ready to train with machine learning algorithms. Therefore, we first transform the original raw text data into structured representations of derived features. These features are referred hereafter as writing style features and include some special features which are not used in the authorship attribution of traditional writings. Because of the casual nature of the Web-based environment, authors are more likely to leave their own *writeprints* in their online conversations (email, chats, forum posts...). Many users try hard to differentiate themselves from others through unique and special online writing style. For example, some authors have the habit of ending the online posts with a smile emoticon “:)” because they think it’s cool. In another case, some people, especially the non-native English speakers, tend to repeatedly use some misspelled words in their sentences, e.g. “absense”, “noticable”, “succesful”... For the Dark Web environment that we consider in this research, there are additional fingerprints that can be taken into accounts. For example, a drug-addicted person may mention a lot about a specific drug name or an active vendor will copy/paste an advertising of his product on multiple Dark Web forums. As a result, with a carefully selected set of features that can represent those fingerprints, the performance of the authorship matching can become reasonably good.

5.4 Validation Phase

We collected posts of 10 active users in different Dark Web forums from the Dark Web archive dataset. We defined the active users as the ones with at least 400 posts and around 6000 words (the contents must be written mostly in English). We denote this set as A. In this validation phase, we divide A into two parts, each of which contains half the posts of each user. We denote those two subsets of A as A_1 and A_2 respectively. We also collect posts of 99 other active users which are used as the fixed negative training dataset for both experiments and denote this set as C. The data in A_1 is used as the positive training data while the data in C provides the negative points which help to shape the classifiers. The desired result is 10 different classifiers for 10 testing users. Specifically, the classifier of each user U is constructed from the training dataset which contains posts of 100 authors (posts of 99 users in set C and posts of user U in A_1). The trained classifiers will be tested against the testing data in the set A_2 using different combinations of parameters. This experiment serves as the validation step which can help to determine the best parameter settings for our model.

5.5 Testing Phase

In the second experiment, for each user of the set A, we find another active account (having at least 400 posts with length around 6000 words) in a different Dark Web forum having the same username and collect all of his forum posts. We denote this set of data as B. The optimal model trained in the validation phase is applied to set B.

6 Discussions

Based on the fact that SVM only performs binary classification task, we apply the N two-way classification model, where N is equal to the number of testing authors. As a result, each SVM classification was applied N times on the testing set of documents. This is known as the one-against-all approach. To predict a new instance of the testing set, we choose the classifier with the largest decision function value. The generated results are in the form of confusion matrices, each matrix's size is $N \times N$. To evaluate the performance of this classification model, we compute the following metrics: recall (R), precision (P) and F_1 measure. F_1 is a popular measure that is frequently used in classification problems, where:

$$F_1 = \frac{2RP}{R + P}$$

For the first validation experiment, we apply the classification model using different SVM kernels. Basically, a kernel is a similarity function which we provide to a machine learning algorithm. It takes two inputs and estimates how similar they are. We carried out the experiment independently for each kernel type. The confusion matrices of all experiments are recorded. The row labels of each matrix are the names of the N classifiers used in the experiments (training authors) while the column labels are the N labeled testing documents. The number in each row (i, j) is the classifying decision value (distance to hyperplane) that the classifier in row i classifies the testing data of author in column j. The label (author) of each test document i is the highest value in column i which is highlighted in red color. After that, using the optimal kernel (which leads to the best accuracy), we apply the model to the testing dataset. The following tables (Tables 1, 2 and 3) show the results of the validation experiments using different kernels. The authors' usernames (labels) are abbreviated in order to protect their personal identity and privacy.

From the above results, we can observe that the linear and polynomial kernels can achieve 10/10 correct classifications. Table 4 shows the F_1 scores of the validation experiments.

In the next step, we carried out the test experiment using linear and polynomial kernels which achieved the perfect classification results for the first

Table 1. Validation experiment result (linear kernel)

	BM(t)	CW(t)	FT(t)	Ka(t)	Ta(t)	Zi(t)	cv(t)	ma(t)	mu(t)	ri(t)
BM(tr)	0.018	−1.154	−2.3	−1.91	−0.541	−0.776	0.777	−1.695	−1.467	−1.554
CW(tr)	−1.519	−0.321	−1.598	−2.091	−0.035	−1.079	−1.821	−2.18	−2.336	−1.858
FT(tr)	−1.507	−0.944	−0.407	−1.235	−1.634	−1.08	−1.272	−0.639	−1.005	−1.084
Ka(tr)	−1.943	−1.953	−1.226	0.144	−2.857	−1.888	−1.961	−0.611	−0.393	−0.91
Ta(tr)	−2.158	−2.541	−3.203	−3.628	1.582	−3.48	−2.795	−3.782	−3.865	−3.817
Zi(tr)	−1.813	−0.932	−1.387	−1.066	−4.126	0.816	−2.009	−1.02	−0.882	−0.722
cv(tr)	−0.927	−0.984	−1.413	−1.372	−0.959	−0.791	1.633	−1.329	−1.199	−1.132
ma(tr)	−1.64	−1.469	−1.082	−1.063	−2.783	−1.314	−1.633	0.531	−0.801	−1.382
mu(tr)	−2.443	−1.95	−1.232	−0.828	−5.082	−1.311	−2.812	−0.411	0.492	−1.1
ri(tr)	−1.624	−1.381	−1.485	−1.093	−1.734	−1.197	−0.709	−1.545	−1.289	0.544

Table 2. Validation experiment result (polynomial kernel)

	BM(t)	CW(t)	FT(t)	Ka(t)	Ta(t)	Zi(t)	cv(t)	ma(t)	mu(t)	ri(t)
BM(tr)	0.103	−1.16	−2.256	−1.85	−0.291	−0.79	0.924	−1.614	−1.423	−1.535
CW(tr)	−1.422	−0.446	−1.519	−1.986	−0.132	−1.134	−1.549	−2.043	−2.15	−1.742
FT(tr)	−1.589	−0.938	−0.377	−1.238	−2.058	−1.082	−1.309	−0.675	−1.007	−1.086
Ka(tr)	−2.083	−2.071	−1.229	0.127	−3.721	−1.918	−2.081	−0.704	−0.431	−0.927
Ta(tr)	−2.059	−2.365	−2.865	−3.157	1.696	−3.043	−2.559	−3.298	−3.318	−3.281
Zi(tr)	−1.912	−0.89	−1.418	−1.109	−5.624	0.868	−2.018	−1.091	−0.898	−0.749
cv(tr)	−0.914	−0.98	−1.374	−1.346	−0.954	−0.832	1.65	−1.296	−1.191	−1.141
ma(tr)	−1.757	−1.518	−1.067	−1.061	−3.709	−1.34	−1.752	0.588	−0.794	−1.424
mu(tr)	−2.703	−2.066	−1.198	−0.851	−7.206	−1.34	−3.192	−0.447	0.476	−1.093
ri(tr)	−1.732	−1.411	−1.478	−1.11	−2.259	−1.208	−0.678	−1.537	−1.285	0.539

Table 3. Validation experiment result (rbf kernel)

	BM(t)	CW(t)	FT(t)	Ka(t)	Ta(t)	Zi(t)	cv(t)	ma(t)	mu(t)	ri(t)
BM(tr)	−0.952	−0.995	−1.105	−1.061	−0.998	−0.986	−0.988	−1.003	−1.018	−1.031
CW(tr)	−1.02	−1.003	−1.001	−1.001	−0.989	−1	−0.989	−0.995	−0.999	−0.995
FT(tr)	−1	−1.019	−0.825	−0.976	−0.987	−1.015	−0.988	−0.971	−1.003	−0.954
Ka(tr)	−1.038	−1.056	−1.019	−0.71	−1.033	−1.062	−1.033	−0.986	−0.91	−0.981
Ta(tr)	−0.997	−1.008	−1.003	−1.003	−0.61	−0.999	−0.952	−0.977	−0.995	−0.979
Zi(tr)	−1.09	−1.091	−1.011	−0.956	−1.082	−0.79	−1.083	−1.068	−1.057	−1.012
cv(tr)	−0.936	−0.982	−1.015	−1	−0.963	−0.997	−0.423	−0.983	−0.995	−0.984
ma(tr)	−0.998	−1.006	−1.008	−1.006	−0.96	−1.008	−0.961	−0.374	−0.954	−0.991
mu(tr)	−1.021	−1.033	−1.069	−0.905	−1.019	−1.067	−1.02	−0.91	−0.66	−1.009
ri(tr)	−1.004	−1.041	−1.002	−0.947	−0.977	−1.036	−0.978	−1.007	−1.02	−0.393

Table 4. Validation experiments F_1 score

Kernel	F_1 **macro**	F_1 **micro**
Linear	1	1
Polynomial	1	1
Rbf	0.765	0.8

Table 5. Test experiment result (linear kernel)

	BM(t)	CW(t)	FT(t)	Ka(t)	Ta(t)	Zi(t)	cv(t)	ma(t)	mu(t)	ri(t)
BM(tr)	−0.073	−0.62	−2.023	−1.841	−1.204	−1.036	−0.328	−1.61	−1.177	−1.952
CW(tr)	−1.678	−0.686	−1.635	−2.254	−3.024	−1.48	−1.831	−2.608	−2.754	−2.58
FT(tr)	−1.39	−0.995	−0.662	−1.144	−1.468	−1.094	−1.14	−0.932	−1.169	−0.903
Ka(tr)	−2.119	−2.084	−1.632	0.164	−1.258	−1.793	−2.333	−0.519	−0.492	−0.649
Ta(tr)	−2.649	−3.012	−2.947	−3.417	−4.152	−3.083	−4.113	−4.027	−4.216	−4.571
Zi(tr)	−1.326	−0.36	−1.64	−1.532	−1.642	−0.063	−0.758	−0.967	−0.559	−0.551
cv(tr)	−1.004	−0.683	−1.585	−1.271	−1.306	−1.04	0.188	−1.273	−1.257	−1.371
ma(tr)	−1.588	−1.353	−1.324	−1.134	−1.411	−1.327	−1.342	−0.327	−0.946	−1.078
mu(tr)	−2.019	−2.043	−1.678	−0.899	−2.096	−1.681	−1.369	−0.616	0.831	−0.372
ri(tr)	−1.703	−1.419	−1.563	−1.011	−1.26	−1.27	−1.042	−1.415	−1.286	0.473

experiments. The results of this testing phase are recorded in Tables 5 and 6. Table 7 shows the F_1 scores of the test experiments.

The results show that there is not much difference between the performance of linear and polynomial kernels. Both can achieve 80% accuracy (8/10 correct classifications) which is a quite good result in authorship attribution domain.

Table 6. Test experiment result (polynomial kernel)

	BM(t)	CW(t)	FT(t)	Ka(t)	Ta(t)	Zi(t)	cv(t)	ma(t)	mu(t)	ri(t)
BM(tr)	−0.049	−0.623	−2.008	−1.801	−1.206	−1.042	−0.421	−1.55	−1.167	−1.824
CW(tr)	−1.546	−0.777	−1.518	−2.072	−2.583	−1.425	−1.778	−2.341	−2.451	−2.363
FT(tr)	−1.429	−0.994	−0.634	−1.151	−1.455	−1.099	−1.138	−0.944	−1.163	−0.925
Ka(tr)	−2.224	−2.162	−1.67	0.176	−1.238	−1.849	−2.28	−0.582	−0.551	−0.723
Ta(tr)	−2.455	−2.724	−2.679	−3.013	−3.532	−2.77	−3.486	−3.436	−3.538	−3.764
Zi(tr)	−1.334	−0.294	−1.701	−1.558	−1.625	0.002	−0.762	−1.013	−0.611	−0.656
cv(tr)	−1.005	−0.715	−1.535	−1.262	−1.315	−1.046	−0.075	−1.259	−1.249	−1.349
ma(tr)	−1.683	−1.382	−1.353	−1.138	−1.45	−1.357	−1.38	−0.303	−0.968	−1.103
mu(tr)	−2.15	−2.158	−1.707	−0.916	−2.139	−1.729	−1.44	−0.648	0.741	−0.447
ri(tr)	−1.779	−1.441	−1.574	−1.026	−1.314	−1.285	−1.077	−1.413	−1.287	0.255

Table 7. Test experiments F_1 score

Kernel	F_1 macro	F_1 micro
Linear	0.747	0.8
Polynomial	0.747	0.8

7 Conclusions and Future Work

In this research, we have built a framework for authorship matching based on features extracted from online messages in Dark Web forums. We have undertaken two separate experiments to evaluate the effectiveness of the proposed method. These experiments have shown that writing style can be used to attribute and correlate authorship between users on multiple Dark Web forums with high accuracy. The results of the experiments show that our method is comparable to all tested state-of-the-art methods. We believe that the proposed framework has the potential to aid the task of tracing secret cyber criminal identities that are hidden under the indexed surface webs.

However, there are some constraints associated with this research. Firstly, we only consider stylometric attributes that work at the character level, word level, sentence level, and document level. There are some other attributes that work at the phrase level, the clause level, and the paragraph level. These attributes, however, are not considered in this study. Given their properties, we can assume that these attributes should provide good results at least for the text summarization task. Secondly, as mentioned in previous chapters, we assume that no text obfuscation attempts are made by the users. The problem can be complicated if the users intentionally alter their writing style to avoid being undercovered by legal parties. Therefore, sophisticated techniques for detecting stylistic deception in written texts need to be integrated for handling such complex scenarios. Last but not least, the size of our test dataset (users who have at least two active accounts on different Dark Web forums) is small due to the limitations of the archived data that we use. We can overcome this problem by creating a data crawler to collect more data directly from Dark Web forums which can potentially help to increase the dimension of our test dataset.

In addition to the limitations of this research that we need to overcome, we have identified several promising research ideas based on the current study. There are two directions that can be considered as our potential future works.

(i) Firstly, as this study only focused on English messages, we plan to include more languages into our future researches.
(ii) Secondly, we will try to focus more on other features that are less related to writing style such as topic, posted date/location, user's forum signature, semantic features and apply this technique to find the correlations of users between dark webs and social networks.

References

1. DailyMail: Teen was hours away from columbine-style massacre at his old school: Ex-student, 19, 'stockpiled weapons and explosives including a 9mm pistol and five pipe bombs in a bid to carry out mass murder, July 2015
2. Abbasi, A., Chen, H.: Applying authorship analysis to arabic web content. In: Kantor, P., Muresan, G., Roberts, F., Zeng, D.D., Wang, F.-Y., Chen, H., Merkle, R.C. (eds.) ISI 2005. LNCS, vol. 3495, pp. 183–197. Springer, Heidelberg (2005). doi:10.1007/11427995_15
3. Zhang, Y., Zeng, S., Fan, L., Dang, Y., Larson, C.A., Chen, H.: Dark web forums portal: searching and analyzing jihadist forums. In: IEEE International Conference on Intelligence and Security Informatics ISI 2009, pp. 71–76. IEEE (2009)
4. Burrows, J.F.: Word-patterns and story-shapes: the statistical analysis of narrative style. Literary Linguist. Comput. **2**(2), 61–70 (1987)
5. Wang, R.Y., Storey, V.C., Firth, C.P.: A framework for analysis of data quality research. IEEE Trans. Knowl. Data Eng. **7**(4), 623–640 (1995)
6. Diederich, J., Kindermann, J., Leopold, E., Paass, G.: Authorship attribution with support vector machines. Appl. Intell. **19**(1–2), 109–123 (2003)
7. Zheng, R., Li, J., Chen, H., Huang, Z.: A framework for authorship identification of online messages: writing-style features and classification techniques. J. Am. Soc. Inf. Sci. Technol. **57**(3), 378–393 (2006)
8. Perito, D., Castelluccia, C., Kaafar, M.A., Manils, P.: How unique and traceable are usernames? In: Fischer-Hübner, S., Hopper, N. (eds.) PETS 2011. LNCS, vol. 6794, pp. 1–17. Springer, Heidelberg (2011). doi:10.1007/978-3-642-22263-4_1
9. Kacmarcik, G., Gamon, M.: Obfuscating document stylometry to preserve author anonymity. In: Proceedings of the COLING/ACL on Main Conference Poster Sessions, pp. 444–451. Association for Computational Linguistics (2006)