

A Message Length Verification of Modern Messaging Systems

Abdulaziz Altamimi^{1*}, Faisal Shaman¹, Abdulrahman Alruban^{1,2}

¹Centre for Security, Communications and Network Research, Plymouth University, Plymouth, UK

²Computer Sciences and Information Technology College, Majmaah University, Al Majma'ah, Saudi Arabia

Email: *Abdulaziz.altamimi@plymouth.ac.uk, Faisal.shaman@plymouth.ac.uk, Abdulrahman.alruban@plymouth.ac.uk

How to cite this paper: Altamimi, A., Shaman, F. and Alruban, A. (2020) A Message Length Verification of Modern Messaging Systems. *Journal of Computer and Communications*, 8, 103-113.

<https://doi.org/10.4236/jcc.2020.83010>

Received: February 25, 2020

Accepted: March 14, 2020

Published: March 17, 2020

Copyright © 2020 by author(s) and

Scientific Research Publishing Inc.

This work is licensed under the Creative

Commons Attribution International

License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Nowadays, there has been a rapid increase in the variety and popularity of messaging systems and social networks. It is imperative to consider the effect and impact of the number of words feature on the verification process for modern messaging systems such as Twitter, Facebook, SMS and Email. Given the volume of text is often a restricted factor (due to the nature of messaging systems), key to this investigation is a better understanding of what length of message is required to improve performance. A large historical dataset containing 50 participants, the four datasets containing a large number of messaging system samples (4539 samples for Facebook, 13,616 for Twitter, 6538 for Email and 106,359 for Text message), the best performance was for Text messages, with an EER of 7.6% if the number of words was more than nine; followed by Email with an EER of 14.9% if the number of words was between 25 to 60; then, Twitter tweets, with an EER of 22.5% if the number of words was less than ten. Finally, the Facebook platform with an EER of 31.9% if the number of words was over 11.

Keywords

Author, Word Length Verification, Social Networks, Messaging Systems

1. Introduction

Around 500 million tweets are sent, and 4.3 billion Facebook messages are posted, every day; in addition, more than 200 million emails are sent and approximately two million new blog posts are created daily, and around 15 billion texts are sent every minute around the globe [1]. Research has shown that it is popular (typically for someone in their 20 s) to utilise multiple messaging systems [2]. For example, a study by [3] reports that 64% of Facebook users also had ac-

counts on Myspace, and LinkedIn shared 42% of its members with Facebook and 32% with Myspace.

However, despite the popularity of messaging systems, they are often found to be the source and target of criminal activities. Messaging systems have become an ideal place for criminals due to their characteristics such as anonymity [4] [5] ease of use and low cost [6]. This leads to a variety of direct and indirect criminal activities, such as sending spam texts to gain personal information [7], grooming children, kidnap, murder, terrorism and violence [8] [9].

A need exists, therefore, to be able to identify the ownership of messages shared on these modern systems. Unfortunately, relying on just the account details to simply verify the author of that account could be misleading because messaging platforms typically do not enforce identity checking, thereby enabling the creation of fake accounts or accounts which are not easily traced back to an individual [10] [11]. Authorship verification is, however, an approach that provides the ability to determine the authenticity of the author through an examination of the message and Message Length Verification is crucial.

The remainder of this paper is organized as follows. Section 2 presents literature review of message length verification. Section 3 presents methodological Approach. Section 4 presents experimental results, and subsection 4.1 presents investigating user level performance. Finally, section 5 highlights the conclusions of this paper.

2. Literature Review of Message Length Verification

Several researchers have examined the effectiveness of length of word feature for authorship authentication and identification with text in the range of 75 to a few hundred words. For instance, the study by [12] developed an instant message intrusion detection system framework in order to test the instant message conversation logs of four users, based on 69 stylometric features, focusing mainly on examining character frequency as a stylometric feature, with some additional stylometric features, including: sentence structure, predefined specific characters, emoticons, and abbreviations analysis. The study was an attempt to analyse 2500 characters, which is 500 words, assuming that (1 word = 5 characters). The naive Bayes classifier was used, and it achieved an accuracy rate of approximately 68%. The results show that uppercase characters, special characters and numbers are distinguishable, and can be used as a form of intrusion detection system. According to [13] identifying and showing these features are the main challenge for authorship identification, since they can contain emoticons, special characters and uppercase or lowercase letters.

In the same context of using limited words, for gender identification, the study by [14] investigated four users; each user had 253 Emails and messages ranged from 50 to 200 words per Email. They used function words, structural, stylistic, gender attribute features and SVM for the classification, and they achieved an accuracy rate of approximately 70.2%. Their approach distinguishes

between male and female authors, and the main finding is that function words provide the most important aspect of discriminating gender.

In general, the performance of different types of long documents achieved an accuracy rate of 70% to more than 90% for 50 - 200 words. In most previous studies on long documents, the minimum number of words was found to be 50 words.

However, with microblogs or social network messaging systems, users can simply post a message as a quick update of their status or the activity they are involved in. Twitter is one of the social networks that places a restriction on the amount of text, which restricts its users to a maximum of 140 characters. The study by [15] investigated authorship identification on Twitter, and collected data from only 12 users, with 120 - 900 tweets per user. The feature set used comprised of Style Markers and Bag-of-Words (BOW). The number of style markers used was 86, and these contained punctuation, long words, part of speech, hyperlinks, and other similar attributes. The BOWs contained all the words that came from the raw data, which was used as a measure when the words appeared more than five times in the whole dataset. SVM was used as a classifier. They found out that Style Markers performed better than BOWs for short text, with an accuracy ranging from 60% to 76.75% for BOWs, and 75.1% to 92.3% for Style Markers.

The study [16] focused on author identification for Email. They investigated nine users of short Emails, with approximately 75 words each and with a range of 174 to 706 Emails per user; SVM was used as the classification engine, which produced around 86.74% accuracy.

3. Methodological Approach

The methodology of measuring the message length performance has been divided into two methods: the first method was used to determine the number of words required for each user on each platform for the four historical datasets, and it has been proposed to base this on the average word and median value for the number of words across the historical dataset. The second method is the verification process.

In the first method, in order to determine and define the number of words for authors required on each platform, the following steps were applied to each author on each platform:

- The average number of words per user on each platform was calculated.
- The first median was used to describe the central tendency of the number of word limits for all users' data by calculating the median for each platform; the reason for using the median is to find out the following limits: the lowest number of words, the average number of words, and the longest number of words for that platform. Once the first median was calculated for all users' average number of words, this value is considered to be the longest word length for that platform.
- The investigation focus is on limited and small words, so the second median

was used to calculate the other lowest, and so the longest words have been ignored.

- The figures were divided into three groups: the first median was used to determine the longest words, the second median was used to determine the smallest number of words, and the third group in the middle was used to calculate the value between these two (the values between the largest and smallest words).

Table 1 below shows the results of statistically splitting the groups of number of words in the experiment based on the average number of words per author on each platform.

The second method involved verification procedures as follows:

- Splitting data into a ratio of 70/30 for train and test, since it has been shown to be the best from among all other splitting.
- The Gradient boosting (GB) classifier was used to test the length of word feature, since it is the first time this classifier was used with this specific feature across four platforms to advance the state of knowledge and enable a better decision-making process.
- Prioritising the features in terms of discriminative information prior to being applied to a standard supervised training methodology, RF was used for identifying only the most relevant features. The RF algorithm deals with this as a two-class classification problem.
- In train and test, each group was trained and tested based on determining the number of words that were given and specified in the first method.

Figure 1 illustrates the process of the methodology, and the experimental approach to the number of words, including the feature for the specified number of words fitted into the classifier in order to class them based on the two-class problem used to verify them.

4. Experimental Results

As shown in **Figure 2**, visualising the total number of words for the population

Table 1. Number of word groups.

Platforms	Group 1	Group 2	Group 3
Twitter	<10	10 - 13	>13
Text message	<5	5 - 9	>9
Facebook	<6	6 - 11	>11
Email	<25	25 - 60	>60

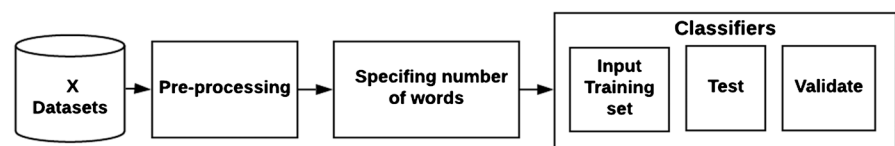


Figure 1. Methodology for the number of word verification.

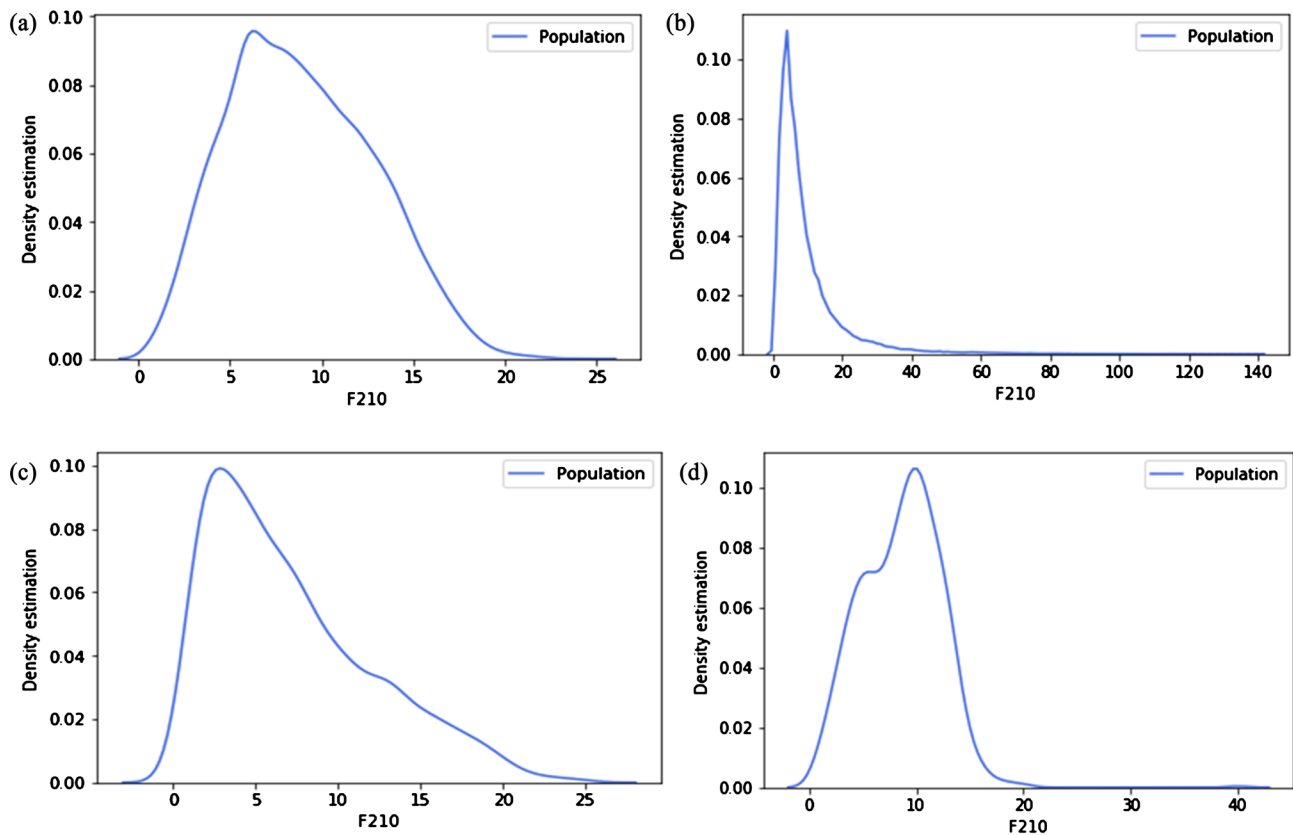


Figure 2. Total number of words for population for Twitter, SMS, Facebook. (a) words for the Twitter; (b) words for the Text message; (c) words the Facebook; (d) words for the Email.

that is on each platform (Twitter, Text message, Facebook and Email) has been determined in order to investigate more about the total number of words for each platform in specific detail, as well as the distribution of word numbers for the population on all platforms in the historical datasets, and the details are presented in **Figure 2** above. **Figure 2** shows the total number of words for the population for Twitter, Text message, Facebook and Email.

It can be seen that in a comprehensive survey of all platforms, the majority of authors on Twitter tend to use approximately less than 10 words, while the same thing occurs on the Text message platform, as authors tend to use approximately 10 words, and the same goes for the Facebook platform, as almost the same range of words of approximately less than 10 words tend to be used. However, for Email the situation is different, as the majority of authors of Email tend to use approximately more than 10 words.

In the case of Twitter, as shown in plot (a), the majority of authors used #words in general that were an average of two to 20 words long in their tweets; however, most authors tend to use approximately less than 10 words. This is expected, as authors have to find a way of being brief and short in their tweet messages using a limited number of words [17].

Similarly, in the case of Text messages, as shown on plot (b), the majority of authors used an average of two to 40 words length for their text messages, and

most authors tended to use approximately less than 10 words; again, this is because authors on Text message have to find a way of being concise and short in their messages. Plot (c) shows that the majority of authors on the Facebook platform used an average of two to 25 words in their posts, and most authors tended to use approximately less than 10 words, this is expected as Facebook messages are usually short in nature [18].

While the majority of authors on the Email platform used words that were an average of two to 20 words long; however, most of them tend to use more than 10 words; however, Emails, on the other hand, allow for a large range of flexibility, and they could vary from just a few words to hundreds of words [19].

Addressing the problem concerning the relative performance of the information that would be necessary to provide reliable verification of an author, requires measuring and characterising the limitations with respect to message length and composition, to ensure reliable author verification decisions. Dozens of experiments were conducted on the historical dataset to examine the message length required to understand and enable reliable author verification decisions. **Table 2** below shows the results for classification performance using the GB classifier, and splitting the data into 70/30 for train/test, and this was used in this experiment for all groups on all platforms.

Table 2 demonstrates the performance of the number of words for four platforms: Twitter, Text message, Facebook and Email. Each platform has been categorised into three groups based on the first proposed method in order to investigate what confidence there is in an author verification decision. On average, the best performance of platforms for the experimental results achieved was for Text messages, with an EER of 7.6% if the number of words was more than nine words; followed by Email with an EER of 14.9% if the number of words was between 25 to 60; then, Twitter tweets, with an EER of 22.5% if the number of words was less than ten. Finally, the worst performance from all four platforms and groups was the Facebook platform with an EER of 31.9% if the number of words was more than 11, and the performance of Facebook across groups did not change significantly.

Table 2. Number of word experimental results.

Platform	Performance EER (%)		
	Group 1	Group 2	Group 3
Twitter	(<10 words) 22.5% (EER)	(10 - 13 words) 25.6% (EER)	(>13 words) 23.9% (EER)
Text message	(<5 words) 10.6% (EER)	(5 - 9 words) 10.02% (EER)	(>9 words) 7.9% (EER)
FB	(<6 words) 28.2% (EER)	(6 - 11 words) 29.5% (EER)	(>11 words) 31.9% (EER)
Email	(<25 words) 15.8% (EER)	(25 - 60 words) 14.9%(EER)	(>60 words) 23.3% (EER)

This is expected in terms of the content of the information, as Facebook messages are short in nature [18]. Another factor that impacted on performance is that Facebook is used for public purposes, and the author is often writing to various different people on a variety of topics, and so uses a varied number of words, which may make it difficult for classifiers to pick up and verify the author. Unlike the Email platform, which is often directed to a person or to a known group of people, or predefined for who will receive these Emails; thus, Facebook showed poor performance even if the number content of the information was more than 11 words. This shows that if the content information on the Facebook platform is less oriented and accurate, or directed to certain people, the performance for verifying the user on Facebook improves for the above reasons.

In contrast, the best performance was for Text messages, as if the content information and number of words was more than nine words, it achieved good performance at 7.9%. This is expected, because Text messages are sent to specific users and are considered private messages on a personal platform—often one to one—and the individual person’s words or writing styles are more familiar for the classifiers. Unlike Twitter or Facebook, the author and texts may be directed to specific people and are not for public use, which suggests that size is less likely to be a determining factor, while the nature of the platform’s use has played a role.

From a different perspective, better results for Text messages, if words (>9), means that the classifier is also supposed to be better for Email, because it is also based on individual use and the user writes in their own style. It can be noticed that Email achieved a good performance of 14.9% if the number of words was between 25 and 60. This case is similar to Text messages, because it is a private platform and message topics are familiar, so authors use their own writing style and words, making it easier for the classifier to verify the user. From another point of view, it is noticed that the Text message platform needs more words to provide more reliable performance, while the Email platform needs between 25 and 60 words to ensure reliable performance. This illustrates that the nature of the platforms may also have an effect on the number of words because Text message has a small capacity; therefore, it needs more words to achieve better performance.

Twitter and Facebook messages did not perform better compared to Text messages and Email. This was expected since these platforms (Facebook and Twitter) are similar in nature regarding publicity, which can make it difficult for the classifier to recognise the writing style of the author. On the other hand, it has been noted that Facebook is also worse than Twitter because the capacity of Twitter is as small, and also most authors may be more accurate in their writing and focus more compared to Facebook, as it has a large space for writing. This is another aspect that may contribute toward the better performance of Twitter compared to Facebook.

In general, it can be stated that on the personal and private platforms such as Email and Text message, the increase in the number of words can be more effective for verifying the author's writing style, and the optimal maximum content on the Email platform may be 60 words to deliver good performance. Unlike Twitter and Facebook, the performance improves if the number of words is less, as shown in **Table 2**, so that the classifiers can find any unique number of words that refer to the author to perform well; in addition, since they are public platforms, the topics are diverse, and the writing style is plain as the author is posting to various people. This section has addressed the problem regarding what length of message is required to provide reliable verification of a platform.

Investigating User Level Performance

The authors in this experiment have been selected since they met the previously mentioned conditions. Firstly, they have at least 20 samples across platforms; secondly, they must have four platforms; thirdly, they must have available samples for the number of words feature specified in each group for each platform. In order to investigate the impact and the effect of the number of individual words across the platforms used, and to investigate if it is possible to verify the author based on his/her number of words. **Table 3** demonstrates the performance of all individual authors across four platforms.

Table 3 shows the performance of authors using the message length features previously defined for each of the four platforms. From this table, it can be observed that the Text message and the Email platform display better performance compared to Twitter and Facebook. It can also be seen that some users, such as Authors 1's EER in Text message was 4.3%; 23.6 for Email; 29.6% for Twitter and 36.3% for Facebook. In this sense, the order of the EER ratio for authors across these platforms is as follows: Text message, Email, Twitter and Facebook, ascending in the sense that the pattern of the author can be determined by the ascending range of relative performance in this order. While some authors, such as Author 3 differs, as their EER was 0.0% for Text message; 10.0% for Email; 22.7% for Twitter and 10.0% for Facebook. Furthermore, it can be noted that the difference in the level of the author's pattern according to the relative performance is as follows: Text message, Email, Facebook and Twitter; in this sense, it has been found that Facebook's performance is better than the performance of Twitter for that author, and since Facebook is similar in performance to the Email platform at 10%, this means that the user pattern is closer and exists on these platforms—Text message, Email, Facebook and finally Twitter—in ascending order. While some authors, such as Author 15 differ, as it can be noted that the pattern can be determined according to this order: Text message, Twitter, Facebook and Email. Therefore, the length of message can provide a reliable verification for some authors across the datasets. The ascending order according to relative performance based on the best to the worst performances of the datasets, the better performance for these four platforms is as follows: Text message

Table 3. All individual classification results for all users with 4 platforms.

SMS #words > 9 (group 3), EER 7.9%				Email #words 25 - 60 (group 2), EER 14.9%			
User	EER (%)	Users	EER (%)	User	EER (%)	Users	EER (%)
1	4.3	10	9.6	1	23.6	10	30
2	2	11	4.2	2	9.9	11	0
3	0	12	2.6	3	10	12	27.5
4	12.6	13	0	4	0	13	16.3
5	5.9	14	14	5	13.1	14	-
6	13.5	15	8.3	6	-	15	30
7	9.5	16	5.7	7	10.6	16	25
8	9.6	17	13.3	8	37.5	17	22
9	16.1	18	5.7	9	39.9	18	12.5

Twitter #words < 10 (group 1), EER 22.5%				Facebook #words < 6 (group 1), EER 28.2%			
User	EER (%)	Users	EER (%)	User	EER (%)	Users	EER (%)
1	29.6	10	25	1	36.3	10	0
2	0	11	19.4	2	-	11	-
3	22.7	12	8.3	3	10	12	35
4	39	13	38.8	4	-	13	21.7
5	30.8	14	32	5	37.1	14	45
6	34.3	15	20.2	6	29.7	15	28.5
7	34.4	16	19.4	7	-	16	40
8	21.5	17	32.6	8	36.1	17	40
9	35	18	-	9	28.3	18	8.3

(more than 9 words with a 7.9% (EER)), Email (between 25 to 60 words with a 14.9% (EER)), Twitter (less than 10 words with a 22.5% (EER)) and finally Facebook (less than 6 words with a 28.2% (EER)).

5. Conclusions

In this research, the number of word feature has been investigated to determine the number of words that would be required to ensure the reliable verification of an author across the four modern datasets. The findings of this research have determined the best/worst message length in the investigation for each platform by determining the relative performance and the best and worst word limit for each platform. For example, on average, the optimal length of messages for the experimental results achieved for Text messages was more than nine words, with an EER of 7.9% and the worst if the number of words was less than five, with an

EER of 10.6%; the optimal length of Facebook posts was less than six words, as the EER was 28.2.8%; then, Twitter tweets, as if the number of words was less than ten, an EER of 22.5% was achieved. Moreover, the Email message investigation achieved the longest number of words compared to the other corpora, as the optimal number of words was between twenty-five and sixty, and an EER of 14.9% was achieved.

The best/worst performance of some authors within each corpus has also been determined (*i.e.* the best author's EER for Email was 0% for *Author 4*, and the worst was *Author 15* with an EER of 30%). The best/worst performance of authors across platforms to gather has also been determined (*i.e.* Author 3's performance across platforms was 0%; 10%; 22.7% and 10% for Text message, Email, Twitter and Facebook respectively). In addition, it was found that the authors' performances were better across platforms when comparing the results in ascending order according to relative performance for these platforms.

Therefore, this investigation has sought to provide a foundation technique for investigators of length of words on platforms to track the footage of an author, and consider the relative performance based on the limit on words for each platform regarding what is required for reliable verification.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- [1] Schultz, J. (2019) Micro Focus Blog. How Much Data Is Created on the Internet Each Day? <https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day>
- [2] Almishari, M., Oguz, E. and Tsudik, G. (2015) Trilateral Large-Scale OSN Account Linkability Study. <https://arxiv.org/abs/1910.09534>
- [3] Chung, C.T., Lin, C.J., Lin, C.H. and Cheng, P.J. (2014) Person Identification between Different Online Social Networks. *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, Volume 1, 94-101. <https://doi.org/10.1109/WI-IAT.2014.21>
- [4] Zheng, R., Li, J., Chen, H. and Huang, Z. (2006) A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques. *Journal of the American Society for Information Science and Technology*, **57**, 378-393. <https://doi.org/10.1002/asi.20316>
- [5] Abbasi, A. and Chen, H. (2008) Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace. *ACM Transactions on Information Systems*, **26**, 7. <https://doi.org/10.1145/1344411.1344413>
- [6] Parameswaran, M. and Whinston, A.B. (2007) Social Computing: An Overview. *Communications of the Association for Information Systems*, **19**, 37. <https://doi.org/10.17705/1CAIS.01937>
- [7] Stringhini, G., Kruegel, C. and Vigna, G. (2010) Detecting Spammers on Social Networks. In: *Proceedings of the 26th Annual Computer Security Applications*

- Conference*, ACM, New York, 1-9. <https://doi.org/10.1145/1920261.1920263>
- [8] Page, R., Barton, D., Unger, J. and Zappavigna, M. (2014) *Researching Language and Social Media*. Routledge, New York. <https://doi.org/10.4324/9781315771786>
- [9] Weir, G.R., Toolan, F. and Smeed, D. (2011) The Threats of Social Networking: Old Wine in New Bottles. *Information Security Technical Report*, **16**, 38-43. <https://doi.org/10.1016/j.istr.2011.09.008>
- [10] Nirkhi, S.M., Dharaskar, R.V. and Thakre, V.M. (2012) Analysis of Online Messages for Identity Tracing in Cybercrime Investigation. *International Conference on Cyber Security, Cyber Warfare and Digital Forensic*, Kuala Lumpur, 26-28 June 2012, 300-305. <https://doi.org/10.1109/CyberSec.2012.6246131>
- [11] Maheswaran, M., Ali, B., Ozguven, H. and Lord, J. (2010) Online Identities and Social Networking. In: *Handbook of Social Network Technologies and Applications*, Springer, Berlin, 241-267. https://doi.org/10.1007/978-1-4419-7142-5_11
- [12] Orebaugh, A. (2006) An Instant Messaging Intrusion Detection System Framework: Using Character Frequency Analysis for Authorship Identification and Validation. *40th Annual IEEE International Carnahan Conferences Security Technology*, Lexington, 16-19 October 2006, 160-172. <https://doi.org/10.1109/CCST.2006.313445>
- [13] Ali, N., Hindi, M. and Yampolskiy, R.V. (2011) Evaluation of Authorship Attribution Software on a Chat Bot Corpus. *23th IEEE International Symposium on Information, Communication and Automation Technologies*, Sarajevo, 27-29 October 2011, 1-6. <https://doi.org/10.1109/ICAT.2011.6102123>
- [14] Corney, M., De Vel, O., Anderson, A. and Mohay, G. (2002) Gender-Preferential Text Mining of e-Mail Discourse. *18th Annual Computer Security Applications Conference*, Las Vegas, 9-13 December 2002, 282-289. <https://doi.org/10.1109/CSAC.2002.1176299>
- [15] Green, R.M. and Sheppard, J.W. (2013) Comparing Frequency- and Style-Based Features for Twitter Author Identification. <https://www.aaii.org/ocs/index.php/FLAIRS/FLAIRS13/paper/viewFile/5917/6043>
- [16] Allison, B. and Guthrie, L. (2008) Authorship Attribution of E-Mail: Comparing Classifiers over a New Corpus for Evaluation. In *LREC 2008*.
- [17] Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H. and Demirbas, M. (2010) Short Text Classification in Twitter to Improve Information Filtering. In: *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, New York, 841-842. <https://doi.org/10.1145/1835449.1835643>
- [18] Hussain, A., Vatrapu, R., Hardt, D. and Jaffari, Z.A. (2014) Social Data Analytics Tool: A Demonstrative Case Study of Methodology and Software. In: *Analyzing Social Media Data and Web Networks*, Palgrave Macmillan, London, 99-118. https://doi.org/10.1057/9781137276773_5
- [19] Li, J.S. (2015) An Investigation of Authorship Authentication in Short Messages from a Social Networking Site. <https://doi.org/10.1109/ICEBE.2014.61>