authorship. The POS tagger has been employed to the corpus to label each word with one of 59 POS tags and after that, used the frequencies of all POS bi-grams that appeared at least three times in the corpus as POS feature set. The study showed the use of idiosyncratic features significantly improved the accuracy rate from 61.7% to 71.8%, using decision tree as the classification method. Compared to the study of [32] as mentioned above, and since both used the email as a platform. It has been found that, although the number of users in the study of [32] more than the number of users studied by [20], the use of lexical and syntactic features in the verification/identification of the email is more effective than the use of the idiosyncratic features, because it is possible that the idiosyncratic features has changed during the writing of the email, for example, (English - British) to (English - American). Having said that, the idiosyncratic features of the word “center” instead of “centre”, a word “center” frequently used in the United States of America while a word “centre” originally used in the United Kingdom. Second, the use of the bayesian classifier seems better than using decision tree specifically in the email platform since decision trees suffer from fragmentation in such cases–particularly if little data was available [33].

There are a few researchers examine effectiveness of stylometry for authorship authentication and identification with text in a range of 75 to a couple of hundreds of words. For instance, Orebaugh [38] developed an instant message intrusion detection system framework in order to test instant message conversation logs of 4 users based on 69 stylometric features and focuses basically on examining character frequency as a stylometric feature and additional stylometric features were used include: sentence structures, predefined specific characters, emoticons, abbreviations analysis. The study tried to analysis of 2500 characters which approximately 500 words, assuming that (1 word = 5 characters). The naive Bayes classifier has been used and achieved an accuracy rate approximately 68%. The results show that uppercase characters, special characters and numbers are distinguishing, and it can be used as a form of intrusion detection system. According to Ali et al [39] identifying and showing these features are the main challenge for authorship identification. Since it has emoticons, special characters and uppercase or lowercase.

In the same context of using limited words, in gender identification which is a branch to the authorship problem, Corney et al [40] investigated 4 users; each user has 253 e-mails and ranged from 50–200 words per e-mail. They used function words, structural, stylistic, gender attributes features and SVM has been used as a classification engine, they achieved accuracy approximately 70.2%. Their approach distinguishes between male and female authors. The main finding of Corney’s outputs, function words provide the most important aspect for discriminating gender. Cheng et al [41] observed that gender identification problem can be treated as a binary classification problem such as class (1) for male and class (0) for female.

Previous studies on long texts especially books, online articles, electronic forums and email, the most commonly stylometric features used for authorship studies were lexical (such as word or character frequency) and syntactic (such as function words or punctuation). One of the most significant findings in previous studies for long documents is that the authorship attribution problem has been significantly influenced by using combination of stylometric feature which combine two or more types of stylometric features. Among various combination features were used, combination of lexical with syntactic can be the best approach to identify authors in long documents and more applicable. Because of the variety of words used in long messages and explain the message in the words used use (“but”, “although”, “at”, etc.).

Significant finding in previous studies for long documents is that authorship attribution has been mainly influenced by the machine-learning paradigm. Among different classification techniques, SVM and Bayesian classifier was regularly used. SVM classifier seems better than Bayesian classifier and decision tree, and Bayesian seems outperform on decision tree. In general, the performance in different type of long documents achieved an accuracy rate of 70% to more than 90% for 50–200 words. Table 1 categorized most studies used in previous studies including techniques, type of features, classification and accuracy.

3.1. Stylometric Features on Short Text

This section focusses upon studies that have sought to specifically use short messages. Short text messages are defined in most previous studies based on approximately 75 words or less (look at the comparison features in the Table 2 in discussion section). Instant messages, SMS messages, social network messages are typical shorter messages unlike other online messages or posting like blog posts or online articles. As it can be clearly shown above in stylometry of long texts, the traditional stylometric features, particularly lexical, syntactic features or combinations of them are more applicable. However, with microblogs or social network messaging systems, users could just post a message as a quick update of their status or activity they are involved in. Twitter is one of the social networks that use a restriction on the amount of text which restricts its user by using only 140 character maximum.

Table 1. The summary of literature review of stylometry searches with long text besides its accuracy

Author	No. of Suspects	Feature Types	Classification type	Accuracy	Goals of study
Zheng et al. 2006	20	Lexical, structural, Syntactic, and content specific	SVM decision tree, and NN	97.69%for SVM, 96.66% for NN and 93.36%for C4.5	identification
Tan, et al. 2010	2	13 Syntactic and 4 lexical	Naïve Bayes	81.98%	Identification
Steyvers, et al.2004	85	Author- topics and topic- word models	SVM	72%,	Topic discovery
Stamatatos 2007	50	Common n- gram	SVM	70%	Identification
Pavele, et al.2009	20	Conjunctions and adverbs	Prediction by partial matching (PPM), and SVM	83-86% for PPM 82.9- 84% for SVM	Identification
Monaco, et al.2013	30	Lexical and syntactic	K-NN	91.5%, EER 8.5	Authentication
Koppel, et al.2004	10	Common words or partial word (n-gram)	SVM	95.7%	Authentication
Iqbal, et al. 2010a	158	Lexical, syntactic, and structural	Bayesian network	80.6%, EER 19.4	Authentication
Corney, et al. 2002	4	structures, stylistic function words and gender- attributes	SVM	70.2%	Gender discovery
Baayen, et al. 2002	8	50 Function words, 8 punctuation	Entropy- weighted linear	88.1%	Identification
Orebaugh 2006	4	Sentence structure, emoticon, and abbreviation. etc.	Naïve Bayes	99.29%	Identification
Howedi,et al. 2014	10	Lexical, structural, Syntactic, and content specific character N-gram	Naïve Bayes and SVM	96%	identification
Ragel et al.(2013)	70	unigrams	cosine similarity and the Euclidean distance	25%	Identification

Therefore, certain stylometric features, as structural features, may not be applicable and effective to work because, users do not have so much control over the content of the post and will indeed potentially need to modify their behaviour in order to conform to the restrictions placed upon the messaging platform.

Layton, et al [16] tested 50 Twitter users with each user having 120 Tweets. They used a 3- gram approach and the Source Code Authorship Profile (SCAP). The study obtained an accuracy of 70%. The users of Twitter can use “#” followed by a tag name to link messages with specific topics. Also, the users of Twitter heavily use “@” followed by a specific user’s name to direct the message to a destination. All these structural contents were counted toward the 140-character limitation; however, all these structural contents were removed by the researchers before applying the SCAP algorithm to allow SCAP to focus on the actual content that the user made and they point out that 120 tweets per user is an important threshold. Furthermore, the SCAP method extends [42] on classification [43]. In Keselj’s study [42], an author profile is described as “a set of length L of the most frequent n-grams with their normalized frequencies.” and an n-gram is an n numbers of characters in contiguous sequence. The profile of an author can be indicated as {(x1; f1), (x2; f2)… (xL; fL)} For i=1…L, where xi indicates to an n-gram and fi indicates to the normalized frequency of xi. For SCAP, the frequency fi of the n-gram xi was not normalized. A profile of an author is defined as the L numbers of n-grams which have the highest frequency, which can be as: {x1, x2….xL}. Throughout the classification process and when an unknown profile was accessible for authorship identification, the author who shared the most n-grams with the unknown profile will be specified as the author of the unknown profile. However, the drawback in their method is that any increasing in messages would not have any positive effect on the accuracy, where their method was questioned in relation to the authorship identification task of not so common messages as the accuracy rate dropped by 27% when data about the discusser’s user information was taken out [44].

Table 2. The summary of literature review of stylometric features on short text besides its accuracy

Author	No. of Suspects	Feature Type	Classification	Accuracy	Goals of study
Layton, et al. 2010	50	Character n-grams	SCAP Algorithms.	70%	Identification
Green, et al. 2013	12	Bag-Of- Words and style markers	SVM	12 users were tested gain 40.5%,	Identification
Allison,et al. 2008	9	Word frequency, 2-grams, 3-grams and stem words	Multimodal Hierarchical SVM	78.46% 87.05% 86.74%	Identification
Zheng, et al. 2006	20	Lexical Syntactic, structural	SVM and C4.5	97.69% and 93.36%	Identification

One of the most significant technique to identify the author in Twitter is to deal with n-gram to create the reference template which containing a contiguous sequence of n items of a particular sequence of text which has the power to collect and distinct the characters of Twitter as long as the limits of characters in Twitter is 140.

Similarly, Green et al [45] studied authorship identification in 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 utilized was 86 and contains punctuation, long word, part of speech, hyperlinks, and other similar attributes. The BOWs contain all the words that come from the raw data which used as a measure when the words appeared more than 5 times in the whole dataset. SVM were used as a classifier and, Sequential Minimal Optimization (SMO) in WEKA was used as algorithm, and a linear kernel has been used as a parameter. 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 drawback is that when the researchers examined the effect of the number of authors found out that the accuracy decreased from 92.3% with two more authors were added to become 40.5% with 12 authors. The reason for the low accuracy rate is that the greater the number of users added the lower the accuracy rate. The increase in the number of users has a significant impact on the parameters process especially with the Style Markers. Because each new user seems to have new patterns in writing style that affects the stability of the overall measurement specified in the balancing parameters. Features of style markers may be suitable to identify authors in small samples of dataset because it is possible to be weak when the number of authors increases [46].

In comparison to the previous study [16], it can be inferred that Layton et al.'s their accuracy dropped to approximately 27%, while Green et al's decreased accuracy was approximately 51%, taking into account that the number of authors at Layton is bigger four times in compared to the number of authors at Green, which indicates that n-gram can play an important role and more effective in Twitter. Because the n-gram has the ability to handle characters and has the ability to distinguish them within text, this leads to user identification.

Ragel et al. [21] focused specifically on authorship detection of SMS to identify authorship using unigrams as features. They stated that the length of the SMS is limited to 140 characters. NUS corpus [47] has been used since it contains more than fifty thousand SMS messages written in English. They used two methods cosine distance as well as Euclidean distance in order to choose a suitable comparison method. 20 authors have been used and more than five hundred SMS each. The first method, a unigram authorship profiling has been used and a distribution of unigrams has been created for each author. Furthermore, the extracted unigram profiles compared against the training set. The author is determined based on similarity comparison between the training and the testing data via the cosine and Euclidean distances. The second method, they tried to determine the effect of the size of the training set. They noticed that when the number of possible authors increases, the decrease in the accuracy is close to linear. They sum up with that ten stacked SMS messages is ideal for the close best accuracy to detect the author of a set of SMS messages using cosine similarity distance metric with unigram method. The drawback is that their results demonstrated a low precision of below 25%, among only 70 SMS users. However, this can reflect the difficulty of obtaining a high precision in authorship identification of SMS users and more appropriate to use another effective technique such as a Part-of-Speech POS combined with n-gram to decrease the number of features.

Allison et al. [18] focused on author identification of email. They investigated 9 users on short emails with approximately 75 words each and with a range of 174 to 706 emails per user, Enron Email corpus has been used and used 2-grams, 3-grams and word frequency measures. SVM was used as the classification engine, which produced around

86.74% accuracy. In comparison with the previous studies mentioned using the e-mail platform for short text, [18] and [40] both utilized SVM as a classifier, Allison achieved 86.74% while Corney et al. [40] achieved 70.2%, Keep in mind that the message length of Allison is 75 words and Corney's message length is between 50-200 words and the differences is only in the number of users and the number of chosen features in stylometry. This indicates to that there is a weakness to determine the best features to be used in stylometry, the best size of the e-mail message, and the appropriate number of users of e-mail platform. In order to be used in the identification process. This leads to discover and to find out the optimization for features mentioned above, because there is no solid base for reliance on during the investigation or exploration.

Koppel et al. [48] and Sanderson et al. [49] investigated 500 words from books and newspaper journalists, respectively. They used an approach "Author Unmasking", the idea of author unmasking is that the differences between two texts from the same author will be reflected in a relatively small number of features. These features can be extracted by use of an author unmasking curve. In Koppel et al.'s experimental, they selected the 250 most frequent words and the collection of 21 English books published in nineteenth century has been used. These books were written by ten different authors. Each book has been chunked into equal sections of at least 500 words without breaking up paragraphs. A linear-kernel SVM has used for cross-validation; they obtain overall accuracy of 95.7%. In Sanderson et al's experimental, they used 50 newspaper journalists, and the number of authors about 50 authors with a minimum of 10,000 words per journalist, they divided the training set into 500 characters per chunk used. A Linear SVM classifier has been used and the accuracy achieved over 90%. They suggested that there is noteworthy aliasing between different-author and same-author performance curves when considering samples of 5,000 words or less and stated that the author unmasking approach is less useful when dealing with short texts. They concluded that measuring of the "depth of difference" between two example sets is a different type than other measures, such as margin width, that could be based on a single highly differentiating feature and it is not appropriate this measure should be applicable to other applications.

Siham and Halim [50] also maintained the idea that the longer the text, the better the identification accuracy will be. The possibility of short documents analysis is difficult to carry out since this type of document is usually characterized by a poor structure and an informal language much less seen in literary texts [23]. For example, the content of the short message that exists in social networks such as Twitter and Facebook is often written without author's conscious; this is due to the fact that the writing practice provided by the platforms does not allow more contents and may be restricted by certain linguistic restrictions, thus, the text message can be unclear language and unorganized format.

In general, the researchers did not agree on a clear vision and a general framework used when the message is being poorly structure and uninformally language built to determine its validity and its suitability when it is being used in the investigation process. For example, the length of the required message, weakness, and strength of stylometric features as well as acceptable format. Having said that, the number of the research that were conducted on short texts is smaller in comparison to the research on long texts. The majority of the studies on short text are associated to digital copies, for instance, emails or social network posts, ranging from 140- character text for Twitter to emails of 75 words. Lexical and syntactic features were commonly used on short texts; SVM can be a common classification method.

Most of the previous studies showed that the effect of lexical and syntactic is an important factor in short text messages for the following reasons: Firstly, short text messages attempts to be a summary and specific to be understood by the other party and try to be inclusive of a few words so the author of the letter takes into account the rich words intended. Secondly, Short text messages contains rules for the sentence to be adopted by adding words such as (but, "and","on","therefore", etc.) As well as punctuation. This plays an important role in the process of syntax the text message because most of the texts contain comprehensive and accurate explanation. Thus, for example, a Twitter message containing 140 characters and the purpose of delivering the text message to the other party in a conceptual and concise. Technically, the most characteristic feature to deal with short messages is n-gram because it plays a large role in determining the characters of letters-words in sentences. As we have seen, most previous studies focused on the n-gram and have achieved a high level of accuracy.

The performance of authorship attribution systems on short texts can be affected by several factors. Stamatatos [19] and [3] investigated the effectiveness of parameters in identification for example, how many authors can be expected to be identified as well as the number of messages that can have an influence on recognition performance. Zheng, et al [3] states that in the authorship analysis task the texts should be attributed to a single or more predefined classes based on the authors which means that these texts are grouped into different classes by their authors.

4. Discussion

This part is divided into two sections. First section discusses the mechanism of identifying authors in long text messages as well as short text messages. Section two discusses the mechanism of connecting authors in multiple platforms.

Part I: Discusses the mechanism of identifying authors in long text messages as well as short text messages.

4.1. Identification of authors for long text messages

Regarding identification of authors on long texts especially books, online articles, electronic forums and email, the highest number of users was used, goes for the study [32] used about 158 users compared to all previous studies using the email as a platform and 200 e-mail for each users, the sample size was less than 500 words, features type were lexical, syntactic, and structural, 292 number of features were used by using the Bayesian network and achieved accuracy rate of 80.6%%, their limitation is that the style variation of the same suspect when he writes may affect his representative model. While the lowest number in the number of users were 2 users for the study [52] using 167 blog posts per user, 170-357.5 words sample size, the features types were 13 syntactic and 4 lexical, 21, No.of features were 21 using Naïve Bayes as a classifier, Their limitation is that each author exhibits differences in the length of their entries in the database in terms of word count and achieved accuracy of 81.98%.

Although the difference between the two studies in the number of users and platforms and both of them used lexical and syntactic, but [32] used the feature of structural, and this seems to increase the rate of accuracy and plays an important role in the email platform even if the number of users are large and this features can be the best features because it reveals the manner users style while using email as a platform, while in the Tan's study, they did not use this feature although the number of users is less. The reason for the absence of this features in the Tan’s study is that most users of blog post do not consider these features as long as users write randomly without any conscious and do not write official text messages such as email. The best performance can be for the study [32] because they deal with a large number of users and even the accuracy of the study [52] was slightly increased by almost 1.38%.

From another side at using the email, it can be noted that (Corney et al.) [40] used the email as a platform identification, the number of users 4, samples each user 253 email, the sample size 50-200 words, the features type were structures, stylistic, function words and gender- attributes, No of features 22, the classifier used SVM and achieved 70.2%. Their limitation they mentioned that there is a need for larger set of gender attributes to be used in order to increase the performance results. Although the study [40] and the study [32] both use email as a platform, and despite their differences between them in number of users, features type and classification. In fact, function word feature in the study (Corney et al.) [40] played a major role in increasing accuracy in Corney’s study, and keep in your consideration Iqbal’s study uses structure feature.

Koppel et al. [20] and Monaco et al. [34] both used the book, the number of users in Koppel, et al.’s study was 10 users, samples each suspect 21 books, the sample size about 500 words per chunk, the type of feature used is common word (n-gram), the number of features 250, the SVM as classification type was used their limitation is that unmasking approach might find more general application, and achieved the accuracy of 95.7%. While in the study [34] 10 books were used, No of users was 30, the sample size was 10,000 words, the features used were lexical and syntactic, the number of features was 228, and the classifier was K-NN, their limitation is that the database is relatively small, and achieved accuracy 91.5%.

In general, the most commonly stylometric features used for author identification were lexical (such as word or character frequency) and syntactic (such as function words or punctuation). One of the most significant findings in previous studies for long documents is that the authorship attribution has been significantly influenced by using combination of stylometric features which combine two or more types of stylometric features, since combination of stylometric features, could help to improve the accuracy of identification, and contains lots of features. The longer the text is, the easier it is to compute stylometric features, which become more reliable as more text is considered. Among various combination features were examined, it seems that the combination of lexical with syntactic have achieved high accuracy in identifying authors in long documents and can be more applicable.

4.2. Identification of authors in the short text messages

The second part of this review has focused on identification of authors on short texts. The number of the research are limited. The majority of the studies on short texts have been applied to emails or social network posts, ranging from 140-character text for Twitter to emails of 75 words. Except for the study [3] ranging 84-346 words, which is a slightly higher than authors’ approach since Zheng’s samples for each suspect was less between 30-92 Email samples. This gives the indicator that increasing the number of suspects' samples for short text message is also an important factor for message size, because the greater the number of author's samples the greater the recognition rate is achieved and vice versa.

From another point of view, Layton et al. [16] and Green, et al. [45], both have used Twitter as a test platform. The aim of their study is identifying users on Twitter, the type of features used in the study [16] was n-gram, No.of users 50, samples for each users 120 tweets, sample size 140 chars max, features type n-grams, no of features was not specified, SCAP algorithms was used as a classifier, their limitation is that the accuracy dropped by 27% when data about the discusser's user information was taken out. Where the study [45] their aims comparing frequency and style-based features for Twitter author identification, they used the feature of Bag-Of-Words and style markers, No of users was 12, samples for each users 120-900 tweets, samples size 140 chars max, No.of features were hundreds features of bag- of-words and 86 style markers, classifier type was SVM, the achieved 40.5%. This gives the indication that the type of selected features play significant role in increasing and decreasing the accuracy of identification. Since, the features of style markers may be suitable to identify authors in small set of samples and it is possible to be weak when the number of authors increases,

On the other hand, in the study of [18] aim authorship attribution of e- mail, they used the features word frequency, 2-grams, 3- grams and stem words, samples for each user is 9 and classifier are multimodal, hierarchical and SVM. The two classifiers multimodal and hierarchical are probabilistic, that they derive an explicit estimate for the probability which a new document be appropriate to each of the likely classes, No. of Features is not specified, sample size was 75, their limitation is that complex linguistic features do not allow for successful discrimination, their accuracy were 78.46% for multimodal, 87.05% for hierarchical and 86.74% for SVM. Unigram feature seems to be outperform to bigrams, and to trigrams as long as some certain stylistic texts are doubtless and since they captured by the longer n–grams and can contain more characteristics and this may contribute to the process of getting closer to the identification.

In general, lexical and syntactic features were also commonly used on short texts. Indeed, the unique structural characteristics of messaging systems can also facilitate authorship identification by using these structure and can be important evidences leads to identification and discriminators, for instance, words of beginning of the sentences, greetings, signatures, quotes, links, and could have more important information details lead to understand more about the author. But in case of use in Twitter might be ineffective since Twitter users write informally and randomly and restricted only by the number of specified words.

Part II: Discusses the mechanism of connecting authors in multiple platforms.

This technique is almost new and there are few studies, and it has been reviewed and discussed extensively previously in the literature studies section. The first work on user linking was conducted by Zafarani et al. [53] who tried to connect users across multiple websites, two methods were suggested, URL of a user profile page which contained the corresponding users name and the natural user's profile which contained another community's username. Liu et al. [54] try to build behaviour similarity model and structure information model, and use multi-objective optimization with missing information to identify linkage across social networks. Afroz et al. [30] try to link users that have multiple accounts within the same forum or blog-based site, linking was based on artificially created accounts of the same user. Almishari et al. [22] try to link Twitter accounts based on very lexical features. However, there is a weakness in all these studies for these reasons:

1. The process of linking users depends mainly on the size of the amount of content of the text message, because often the goal of social networking sites depends heavily on messaging and text messages, even with sites that offer video and image services, for example, the comments in the YouTube video is text, Snapchat provides text messaging service, most social networking sites, email and SMS all provide text, thus optimization text with the number of words with each platform and with multiple platforms were not available in all above studies.
2. The nature of the stylometric features for all platforms need be adapted to each other and optimized with stylometric features types with all platforms. For example, the features of stylometry associated with Facebook, Twitter, or the extent of correlation features with each other, alternatively, this also makes it adaptable to the volume of text messages received in order to carry out the identification process and this is also not available in the above studies. According to Goga et al [55]; Jain et al [56] their work used attributes without analyzing their properties and their limits to match profiles in practice and thus they use attributes with low availability which can only match a small portion of profiles across a small number of social networks and is likely to give many false matches in practice.

5. Conclusion

Several methods and systems have been proposed for solving the problem of author identification of short texts, it is still not clear what the volume of message needs to be for reliable and confident verification/identification. The drawback of the previous studies is that they have only been focused upon a single platform, yielding incomplete data and insufficient reliability of recognition when dealing with multiple messaging systems, because most platforms differ from each other, whether technically or linguistically and there are several differences in most multiple platforms on different aspects, for example, in modality, Twitter is public platform in nature, while SMS mostly is used for exchanging private text messages. The most significant aspect, linguistically, the post/tweets text message is not necessarily restricted by caution or fear of people such as SMS messages since they are public and the users can easily post and tweet and can hide themselves without any cost, while the SMS has to have or buy a SIM to start anonymity and they takes into account their anonymity and caution because SMS text message has to be addressed to the user and take the linguistic reserve and consideration about it, so he/she cannot be denied in most cases because he/she is the one who places the address of the message to a practically user or specific group of people.

Finally, the drawback of most previous studies is that they have used multi-objective in order to optimize with only with missing information, yielding unsatisfactory data and inaccurate data to achieve the flexibility and the reliability to deal with these platforms. Therefore, there is a lack of understanding about to what extent these stylometric features are actually portable between messaging system platforms. In addition, lexical and syntactic using n-gram is being the most likely candidates to apply in multiple platforms. Indeed, analysis of lexical and syntactic features across different platforms is important in order that the resulting profile is useful across multiple platforms.

6. References

[1] Marques, O. (2016). Social Networks. In Innovative Technologies in Everyday Life (pp. 31- 44). Springer International Publishing.

[2] The Guardian (2014, January 12). OMG! Number of UK text messages falls for first time. Retrieved 2017, from https://www.theguardian.com/technology/2014/jan/13/number-text-messages-sent-britain- falls-first-time, (Access Date: 19 February, 2019).

[3] 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(3), 378-393.

[4] Abbasi, A., and Chen, H. (2008).Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Transactions on Information Systems (TOIS), 26(2), 7.

[5] Parameswaran, M., and Whinston, A. B. (2007). Social computing: An overview. Communications of the Association for Information Systems, 19(1), 37.

[6] Stringhini, G., Kruegel, C., and Vigna, G. (2010, December). Detecting spammers on social networks. In Proceedings of the 26th annual computer security applications conference (pp. 1-9). ACM.

[7] Page, R., Barton, D., Unger, J., and Zappavigna, M. (2014). Researching Language and Social Media. New York: Routledge.

[8] Weir, G. R., Toolan, F., and Smeed, D. (2011). The threats of social networking: Old wine in new bottles. Information security technical report, 16(2), 38-43.

[9] Maheswaran, M., Ali, B., Ozguven, H., and Lord, J. (2010). Online identities and social networking. In Handbook of Social Network Technologies and Applications (pp. 241-267). Springer US.

[10] Nirkhi, S. M., Dharaskar, R. V., and Thakare, V. M. (2015). Authorship Identification using Generalized Features and Analysis of Computational Method. Transactions on Machine Learning and Artificial Intelligence, 3(2), 41.

[11] Nirkhi, S. M., Dharaskar, R. V., and Thakre, V. M. (2012, June). Analysis of online messages for identity tracing in cybercrime investigation. In Cyber Security, Cyber Warfare and Digital Forensic (CyberSec), 2012 International Conference on (pp. 300-305). IEEE.

[12] Abbasi, A., and Chen, H. (2005).Applying authorship analysis to extremist-group web forum messages. Intelligent Systems, IEEE, 20(5), 67-75.

[13] Alotaibi, S., S. Furnell, and N. Clarke, Transparent authentication systems for mobile device security: A review. In the 10th International Conference for Internet T technology and Secured Transactions (ICITST) (pp.406-413). IEEE. 2015.

[14] Alotaibi, S. and Alruban, A. (2017), Systematic Literature Review of Behavioural Profiling for Smartphone Security: Challenges and Open Problems.

[15] Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., and Woodard, D. (2018). Surveying stylometry techniques and applications. ACM Computing Surveys (CSUR), 50(6), 86.

[16] Layton, R., Watters, P., and Dazeley, R. (2010, July).Authorship attribution for twitter in 140 characters or less. In Cybercrime and Trustworthy Computing Workshop (CTC), 2010 Second (pp. 1-8). IEEE.

[17] Fridman, A., Stolerman, A., Acharya, S., Brennan, P., Juola, P., Greenstadt, R., and Kam, M. (2013). Decision fusion for multimodal active authentication. IT Professional, 15(4), 29-33.

[18] Allison, B., and Guthrie, L. (2008, May). Authorship Attribution of E-Mail: Comparing Classifiers over a New Corpus for Evaluation. In LREC.

[19] Stamatatos, E. (2007, September). Author identification using imbalanced and limited training texts. IN Database and Expert Systems Applications, 2007.DEXA'07. 18th International Workshop on (pp. 237-241). IEEE.

[20] Koppel, M., Schler, J., and Argamon, S. (2013). Authorship Attribution: What's Easy and What's Hard.

[21] Ragel, R., Herath, P., and Senanayake, U. (2013, December). Authorship detection of SMS messages using unigrams. In Industrial and Information Systems (ICIIS), 2013 8th IEEE International Conference on (pp. 387-392). IEEE.

[22] Almishari, M., Kaafar, D., Oguz, E., and Tsudik, G. (2014 a, November). Stylometric Linkability of Tweets. In Proceedings of the 13th Workshop on Privacy in the Electronic Society (pp. 205-208).ACM.

[23] Brocardo, M., Traore, I., Saad, S., and Woungang, I. (2014). Verifying Online User Identity using Stylometric Analysis for Short Messages. Journal of Networks, 9(12), 3347-3355.

[24] Juola, Patrick (2006). "Authorship Attribution". Foundations and Trends in Information Retrieval.

[25] Mariappan, P., Padhmavathi, B., and Teja, T. S. (2016). Digital Forensic and Machin Learning. In Combating Security Breaches and Criminal Activity in the Digital Sphere (pp. 141-156). IGI Global.

[26] Khanum, M., Mahboob, T., Imtiaz, W., Ghafoor, H. A., and Sehar, R. (2015). A Survey on Unsupervised Machine Learning Algorithms for Automation, Classification and Maintenance. International Journal of Computer Applications, 119(13).

[27] Li, J., Zheng, R., and Chen, H. (2006). From fingerprint to writeprint. Communications of the ACM, 49(4), 76-82.

[28] Baayen, H., van Halteren, H., Neijt, A., and Tweedie, F. (2002, March). An experiment in authorship attribution.In 6th JADT (pp. 29-37).

[29] Overdorf, R., Dutko, T., and Greenstadt, R. (2014). Blogs and Twitter Feeds: A Stylometric Environmental Impact Study.

[30] Afroz, S., Islam, A. C., Stolerman, A., Greenstadt, R., and McCoy, D. (2014, May). Doppelgänger finder: Taking stylometry to the underground. In Security and Privacy (SP), 2014 IEEE Symposium on (pp. 212-226). IEEE.

[31] Narayanan, A., Paskov, H., Gong, N. Z., Bethencourt, J., Stefanov, E., Shin, E. C. R., and Song, D. (2012). On the feasibility of internet-scale author identification. In Security and Privacy (SP), 2012 IEEE Symposium on (pp. 300-314). IEEE.

[32] Iqbal, F., Khan, L. A., Fung, B., and Debbabi, M. (2010a). E-mail authorship verification for forensic investigation. In Proceedings of the 2010 ACM Symposium on Applied computing (pp. 1591-1598).ACM.

[33] Caragea, D. (2004). Learning classifiers from distributed, semantically heterogeneous, autonomous data sources (Doctoral dissertation, Iowa State University).

[34] Monaco, J. V., Stewart, J. C., Cha, S. H., and Tappert, C. C. (2013, September). Behavioral biometric verification of student identity in online course assessment and authentication of authors in literary works. In Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on (pp. 1-8). IEEE.

[35] Koppel, M., and Schler, J. (2004, July).Authorship verification as a one-class classification problem. In Proceedings of the twenty-first international conference on Machine learning (p. 62).ACM.

[36] Vorobeva, A. A. (2016, April). Examining the performance of classification algorithms for imbalanced data sets in web author identification. In Open Innovations Association and Seminar on Information Security and Protection of Information Technology (FRUCT-ISPIT), 2016 18th Conference of (pp. 385-390). IEEE.

[37] Koppel, M., and Schler, J. (2003, August). Exploiting stylistic idiosyncrasies for authorship attribution. In Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis (Vol. 69, p. 72).

[38] Orebaugh, A. (2006, October). An Instant Messaging Intrusion Detection System Framework: Using character frequency analysis for authorship identification and validation. In Carnahan Conferences Security Technology, Proceedings 2006 40th Annual IEEE International (pp. 160-172).IEEE.

[39] Ali, N., Hindi, M., and Yampolskiy, R. V. (2011, October). Evaluation of authorship attribution software on a Chat bot corpus. In Information, Communication and Automation Technologies (ICAT), 2011 XXIII International Symposium on (pp. 1-6). IEEE.

[40] Corney, M., De Vel, O., Anderson, A., and Mohay, G. (2002). Gender-preferential text mining of e-mail discourse. In Computer Security Applications Conference, 2002. Proceedings. 18th Annual (pp. 282-289). IEEE.

[41] Cheng, N., Chen, X., Chandramouli, R., and Subbalakshmi, K. P. (2009, March). Gender identification from e-mails. In Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on (pp. 154-158). IEEE.

[42] Keselj, V., Peng, F., Cercone, N., and Thomas, C. (2003, August). N-gram-based author profiles for authorship attribution. In Proceedings of the conference pacific association for computational linguistics, PACLING (Vol. 3, pp. 255-264).

[43] Frantzeskou, G., Stamatatos, E., Gritzalis, S., Chaski, C. E., and Howald, B. S. (2007). Identifying authorship by

byte-level n-grams: The source code author profile (scap) method. International Journal of Digital Evidence, 6(1), 1-18.

[44] Rappoport, R. S. O. T. A., and Koppel, M. (2013). Authorship attribution of micro-messages.

[45] Green, R and Sheppard, J. (2013) “Comparing Frequency- and Style-Based Features for Twitter Author Identification,” Proc. Twenty-Sixth International Florida ArtificialIntelligenceResearchSocietyConference, 2013, https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS13/paper/viewFile/5917/6043, (Access Date: 19 February, 2019).

[46] MacLeod, N., and Grant, T. (2012). Whose Tweet? Authorship analysis of micro-blogs and other short-form messages.210-224.

[47] Chen, H. (2011). Dark web: Exploring and data mining the dark side of the web (Vol. 30). Springer Science and Business Media.

[48] Koppel, M., Schler, J., and Bonchek-Dokow, E. (2007). Measuring differentiability: Unmasking pseudonymous Authors, Journal of Machine Learning Research, 8, 1261-1276.

[49] Sanderson, C., and Guenter, S. (2006, July). Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (pp. 482-491). Association for Computational Linguistics.

[50] Siham, O. and Halim, S. (2012). Authorship Attribution of Ancient Texts Written by Ten Arabic Travelers Using a SMO-SVM Classifier, the 2nd International Conference on Communications and Information Technology (ICCIT): Digital Information Management, Hammamey, 44-47.

[51] Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology, 60(3), 538-556.

[52] Tan, R. H. R., and Tsai, F. S. (2010, October). Authorship identification for online text. In Cyberworlds (CW), 2010 International Conference on (pp. 155-162). IEEE.

[53] Zafarani, R., and Liu, H. (2009). Connecting Corresponding Identities across Communities. ICWSM, 9, 354-357.

[54] Liu, S., Wang, S., Zhu, F., Zhang, J., and Krishnan, R. (2014, June). Hydra: Large-scale social identity linkage via heterogeneous behavior modeling. In Proceedings of the 2014 ACM SIGMOD international conference on Management of data (pp. 51-62). ACM.

[55] Goga, O., Loiseau, P., Sommer, R., Teixeira, R., and Gummadi, K. P. (2015, August). On the reliability of profile matching across large online social networks. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1799-1808). ACM.

[56] Jain, P., Kumaraguru, P., and Joshi, A. (2015, August). Other times, other values: leveraging attribute history to link user profiles across online social networks. In Proceedings of the 26th ACM Conference on Hypertext and Social Media (pp. 247-255). ACM.