International Journal of English Linguistics; Vol. 9, No. 5; 2019
ISSN 1923-869X E-ISSN 1923-8703
Published by Canadian Center of Science and Education

# Towards a Linguistic Stylometric Model for the Authorship Detection in Cybercrime Investigations

Abdulfattah Omar[1&2] & Aldawsari Bader Deraan[3]

[1] Department of English, College of Science & Humanities, Prince Sattam Bin Abdulaziz University, Saudi Arabia

[2] Department of English, Faculty of Arts, Port Said University, Egypt

[3] Department of English, College of Arts & Science, Wadi Aldawaser, Prince Sattam Bin Abdulaziz University, Saudi Arabia

Correspondence: Abdulfattah Omar, Department of English, College of Science & Humanities, Prince Sattam Bin Abdulaziz University, Al-Kharj, Riyadh, 11942, Kingdom of Saudi Arabia. E-mail: a.abdelfattah@psau.edu.sa

Received: June 18, 2019 Accepted: July 17, 2019 Online Published: August 26, 2019

doi:10.5539/ijel.v9n5p182 URL: https://doi.org/10.5539/ijel.v9n5p182

**Abstract**

This study proposes an integrated framework that considers letter-pair frequencies/combinations along with the lexical features of documents as a means to identifying the authorship of short texts posted anonymously on social media. Taking a quantitative morpho-lexical approach, this study tests the hypothesis that letter information, or mapping, can identify unique stylistic features. As such, stable word combinations and morphological patterns can be used successfully for authorship detection in relation to very short texts. This method offers significant potential in the fight against online hate speech, which is often posted anonymously and where authorship is difficult to identify. The data analyzed is from a corpus of 12,240 tweets derived from 87 Twitter accounts. A self-organizing map (SOM) model was used to classify input patterns in the tweets that shared common features. Tweets grouped in a particular class displayed features that suggested they were written by a particular author. The results indicate that the accuracy of classification according to the proposed system was around 76%. Up to 22% of this accuracy was lost, however, when only distinctive words were used and 26% was lost when the classification procedure was based solely on letter combinations and morphological patterns. The integration of letter-pairs and morphological patterns had the advantage of improving accuracy when determining the author of a given tweet. This indicates that the integration of different linguistic variables into an integrated system leads to better performance in classifying very short texts. It is also clear that the use of a self-organizing map (SOM) led to better clustering performance because of its capacity to integrate two different linguistic levels for each author profile.

**Keywords:** authorship detection, forensic linguistics, morphological patterns, lexical features, letter-pair frequencies, self-organizing map (SOM)

## 1. Introduction

With the world-wide impact of computer and internet services on modern life, unprecedented problems and crimes have come to the surface with many negative repercussions. These problems and crimes have come to represent the dark side of the Internet (Davies, Francis, & Jupp, 2016; Sutton & Mann, 1998; Wall, 2003). In social media applications, for instance, unusual illegal acts are committed in ways that pose real threats to personal safety in the use of these applications. A number of authorship identification techniques have been developed with the purpose of detecting the real identities of cybercriminals. However, challenges relating to the practical applications of authorship detection remain. One of these challenges is how to identify authors of very short texts, especially in social media applications. Practically, applications that use conventional classification methods based on the lexical and/or structural properties of a text do not usually yield reliable results in the authorship detection of very short texts. Forensic text types are usually very short and have minimal linguistic features. It is therefore difficult for forensic linguists to develop robust evidence as to authorship due to the lack of sufficient linguistic data available to them.

As a means of addressing this problem, this study suggests that a quantitative morpho-lexical approach, which considers the two main variables of letter-pair frequencies and distinctive words and phrases, offers better authorship detection. I propose that authors usually have habits that are reflected unconsciously in their use of letters and words. As such, analysis of an author's style can readily be carried out through the detection of stable word combinations in a given corpus (Brena, 2011; Makagonov, Espinoza, & Sidorov, 2011). The study of letter-pair frequencies can thus be useful in recognizing the real author of a disputed text. Individuals have distinctive ways of writing that are reflected in their use of letters and this acts as a code or fingerprint by which an author can be revealed. The problem with this approach, however, is that there are different variables, which are difficult for conventional cluster analysis to handle. One way of solving this problem is the use of the self-organizing map (SOM) model due to its effectiveness in processing different variables simultaneously.

This study is based on a corpus of selected tweets on the removal of Confederate monuments in the United States in August 2017. The United States has over seven hundred monuments across the country dedicated to the Confederate soldiers and leaders of the American Civil War who revolted after the US government sought to abolish slavery. In 2015, a number of local governments in the United States decided to remove these monuments because of their connotations of white supremacy and racism. In August 2017, however, a white nationalist rally in Virginia brought renewed attention to the hundreds of Confederate monuments around the country (Holland, 2017; Kenning, 2017). Supporters of Confederate symbols were not happy with the planned removal of these Confederate monuments. They considered these monuments to be a part of US history generally and something of which the great majority of Americans should be proud (Landrieu, 2018; Savage, 2017).

Energized by the violent rioting of nationalists and conservatives in Virginia, counter-protesters, called for the immediate removal of Confederate monuments and statues as symbols of racism and oppression. Some of these counter-protesters did not even wait for local officials to act, toppling Confederate monuments by themselves in Durham, North Carolina, and several other American cities. The political battles and controversial debates over the issue brought a flood of reaction on social media platforms, especially after the US President, Donald Trump, commented on the events on Twitter. He posted three tweets defending the Confederate monuments and describing their removal as "a foolish act". He wrote:

- "Sad to see the history and culture of our great country being ripped apart with the removal of our beautiful statues and monuments. You …
- … can't change history, but you can learn from it. Robert E Lee, Stonewall Jackson—who's next, Washington, Jefferson? So foolish! Also …
- … the beauty that is being taken out of our cities, towns and parks will be greatly missed and never able to be comparably replaced!"

According to commentators, Trump's tweets promoted division and fueled discourses of racism and hatred among social media users (Nossel, 2017; Stolberg & Rosenthal, 2017). Furthermore, many observers linked this online hate speech to real-life incidents. Given this context, it can be seen how this topic provides a good opportunity to extract real-life data for addressing a significant issue with modern social media platforms, such as Twitter—the use of hate-speech and the promotion of racial hatred and violence, which can often amount to criminal behavior. In our case, Twitter was used as an experimental case for testing a new authorship detection model based on the integration of letter-pair combinations along with morphological and lexical features. As such, this study investigates whether the authorship of very short texts can be detected using only linguistic stylometry.

## 2. Literature Review

Recent years have witnessed increasing rates of crime associated with the use of social media networks. These criminal behaviors have included offensive language, hate messages, and even the spreading of violence and terrorism. It may be true to say that rather than being platforms for social interaction, many social media networks have instead become effective tools for posting abuse and sending message containing unpleasant or embarrassing information about others. Criminals use them to encourage hate crimes (targeting people on the basis of such things as religion, race, and sexual orientation) and terrorists to spread propaganda and inspire people from around the world to commit terrorist acts. A major reason for the increase of crimes of this kind is the anonymous nature of social networking or what can be described as their general potential for anonymity. Different social media applications and websites, including Facebook and Twitter, enable cyberbullies to send anonymous and destructive messages to others that can cause harm and even lead to suicide. It has even been revealed that the anonymous nature of Facebook and Twitter has been taken advantage of to influence users'

choices, spending habits, and even political decisions (e.g., Russian intervention in the 2016 US elections). Although companies are constantly developing ways to deter and remove abusive posts, the damage done by such posts usually remains. At the heart of this pseudo-anonymity, different social media channels provide rare opportunities for attackers and cybercriminals to abuse others (by posting and sending verbal abuse and threats; spreading false news and information, etc.) and remain shielded from responsibility for their postings.

Although some may argue for the desirability of anonymous communications in public discourse, the consequences of such anonymity on social stability should be considered too. Some may use fake characters in order to create social unrest and shape public understanding in particular ways. For example, Timberg and Harwell (2018) argue that following the Parkland high school shootings in Florida, thousands of anonymous posts about the attack sought to push false information about one of America's deadliest school shootings. The postings gave false explanations about the massacre and even convinced many followers that the shooter was an active member of a white-supremacist group, which had negative implications for social integrity. The idea that different social media networks have become a potent tool of abuse and deception has made it imperative to address anonymity on the internet and think about novel and effective ways of dealing with this new kind of authorship problems. In spite of the development of different approaches for authorship detection, results in relation to applications where the content is limited remain inconsistent. This applies to both linguistic and non-linguistic approaches to the problem. This study is limited to the investigation of linguistic approaches and it is mainly concerned with addressing the problem of authorship detection using only linguistic methods.

The literature suggests that language has always been a key element in the criminal investigation of authorship detection cases (Coulthard & Johnson, 2010, 2013; Craig, 2004; Schreibman, Siemens, & Unsworth, 2004; Solan & Tiersma, 2012). Although the idea of using linguistic knowledge and methods for determining the authorship of texts is very old, the effectiveness of linguistic analysis became increasingly recognized by both researchers and investigation bodies in the second half of the twentieth century and the development of forensic linguistics (Chaski, 2012; Solan & Tiersma, 2012). This term was first coined by Jan Svartvik in 1968. Svartvik was a linguistic expert whose work contributed to highlighting the impact of linguistics on criminal investigations and on legal activities and procedures (Coulthard & Johnson, 2013). Forensic linguistics may generally be described as the application of linguistic knowledge, methods, and systems to legal settings. It tends to offer a careful and systematic analysis of language that can be used by various professionals, including lawyers, judges, and jury members, in evaluating questions of guilt and innocence in ways that serve justice and help to find out the truth about crimes (Solan & Tiersma, 2012). With the development of computational methods, forensic linguistic approaches have become more reliable and today it is considered "a well-established, internationally recognized independent discipline of study" (Coulthard & Johnson, 2010, p. 5).

In authorship detection applications, forensic linguistics is generally based on the notion of a linguistic fingerprint, which is defined as the process of collecting linguistic data and features that stamp a speaker/writer as unique (Olsson, 2008, 2009). The assumption is that people use language differently, and that this difference between people can be observed just as easily and surely as a fingerprint. To do this, forensic linguistics usually adopts quantitative and statistical methods to investigate the linguistic level/s chosen by the researcher. The majority of these quantitative or statistical linguistic approaches, known as stylometric approaches, are primarily based on statistical investigation of the lexical, syntactic, and/or structural features of social media contents. This has proved unsuccessful in detecting the possible authors of offensive content. This may be attributed to the fact that the language of online social media is usually "highly unstructured, informal, and often misspelled" and therefore unique (Chen, Zhou, Zhu, & Xu, 2012, p. 71). Similarly, Ostrowski (2014) argues that the peculiar nature of social media language, being unorganized and characterized by extensive use of abbreviations, makes it difficult for algorithms based on exploring and investigating only the linguistic and stylistic properties of contents to identify possible authors of disputed texts.

Another problem that is associated with conventional stylometric approaches is that words are represented in the form of single words or n-grams (known as the bag-of-words model) using a vector space model to measure similarity between documents in a given corpus. One major problem with this lexical semantic approach is that it ignores the syntax and contextual meaning of texts. Given the shortness of texts in social media, the lexical frequencies encountered are far too low and cluster analysis will generate spurious results. This leads to sparsity problems, which have negative implications for results based on the frequency of lexical types. Authorship detection based only on single words is therefore unreliable. To sum up, with the anonymous nature of modern internet applications and the tendency of users to use very short texts for illegal purposes, conventional or vocabulary-based clustering methods are neither appropriate nor reliable.

In light of the limitations of lexical and structural analyses of such texts, this study proposes an integrated

quantitative morpho-lexical approach in their place. This approach considers the use of letter-pair frequencies, along with the distinctive lexical features of texts, to build a hierarchical cluster analysis. Successfully grouping similar texts together can help in authorship identification. In traditional applications, documents are represented using single words only—the rationale being that each writer has an identifiable fingerprint that can be detected from the use of letters and that the number of possible variables (i.e., pairs) is quite small and the frequencies are correspondingly enhanced (Moisl, 2009). Furthermore, familiar patterns, as reflected in the use of letter combinations, can be more easily identified. In this way, it is supposed that the use of letter-pair frequencies is appropriate for the nature of the data (very short social media texts). The research question centers on the effectiveness of the use of letter-pair frequencies in supporting clustering performance and improving the authorship identification of anonymous users of social media networks.

## 3. Methods

In order to address the problem of authorship detection of very short texts, this study adopts a quantitative morpho-lexical approach. This is an integrated framework that considers both the quantitative morphological and lexical properties of texts. Although different authorship recognition systems are based on the study of the lexical properties of texts, the use of morphological information is not widely used. Taken together, morphological and lexical properties, or what can be described as stylometric information, are thought to be effective clues in identifying the author/s of given texts. Quantitative morphological and lexical analyses come under the general study of quantitative linguistics—an umbrella term that is concerned with the study of the quantitative properties of linguistic elements with the purpose of understanding and explaining different linguistic phenomena and structures. Although the quantitative study of language dates back to the 19th century, it was only in the closing years of the 20th century that its adoption became widespread. Over the past two decades quantitative linguistic methods have been widely used to address varied problems: natural language processing; machine translation; human intelligence; text classification; authorship detection; and information retrieval. Morphological and lexical analyses have been crucial to such applications.

The proposed technique is carried out in two stages. In the first stage, the morphological and lexical information are extracted from the dataset (in this case tweets) and graphically represented. In other words, quantitative methods are used to capture all the distinctive morphological and lexical properties of the corpus and use them as inputs with the purpose of building a structural (graphical) representation. This representation can then be used to better understand the morphological patterns displayed and the ways the words are built. In our case, the proposed hypothesis is that quantitative morpho-lexical methods are useful for identifying the distinctive morphological, lexical features, and authors' writing style, allowing us to assign texts to their authors. In other words, morpho-lexical analysis based on quantitative methods is useful in finding out authors' categories and thus solving the problems of unknown or controversial authors.

In the second stage, automatic text classification (ATC) methods are used to group similar texts together. The goal of ATC systems is to create clusters that are internally coherent, but clearly different from each other. In authorship attribution/recognition applications and tasks, members of each cluster or category are assumed to be written by the same author. For classification purposes, the self-organizing map (SOM) model is used. The model was first developed by Teuvo Kohonen in 1982 and has become one of the most popular neural network and data-dimensionality models. The function of the SOM is to process unsupervised datasets in a simple way, taking into account the neuron neighborhood, to reveal the similarity between high dimensional data before mapping this data onto a low dimensional map, while retaining the distinctive features of the original datasets (Kohonen, 1982). In SOMs, the vectors, called neurons or nodes, are arranged in a single, usually 2-dimensional, grid. This represents the input layer. Neurons in the input layer then lead out of the grid and after multiple iterations successful neurons form areas with a high density of data points reflecting the underlying clusters in the data (Kohonen, 1990, 1995, 2012).

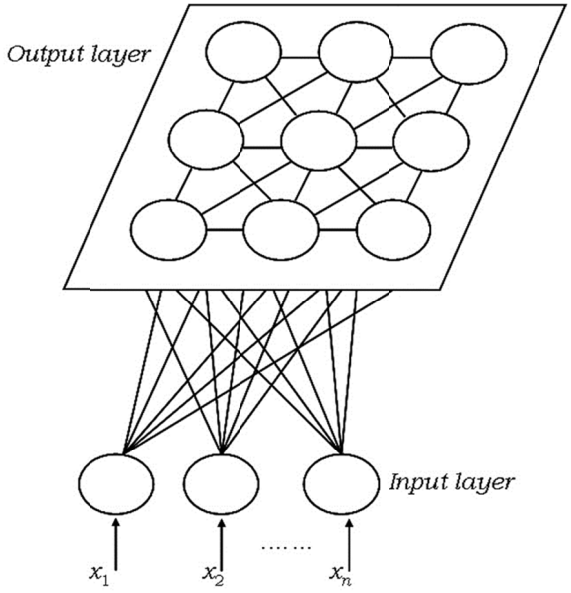


Figure 1. The way SOMs work

Juntunen, Liukkonen, Lehtola, and Hiltunen (2013) argue that the SOM model has a number of advantages over other multivariate approaches such as factor analysis and principal component analysis (PCA). They explain that the SOM model is more effective in dealing with noisy and irregular data and providing more informative interpretations and structures of data with multiple variables. In this way, they assert that it is more visually and easily understandable. The SOM model is effective in enhancing clustering performance and results as it uses both the lexical properties and the relationship between letters (the letter combinations) in the input documents (Johnsson, 2012; Liu, Liu, & Wang, 2012).

In spite of the effectiveness of the SOM model in reducing and visualizing high dimensional data space, as well as retaining the most distinctive features of the inputs, with positive impacts on clustering performance and accuracy, some problems and limitations have been seen when it has been adopted in some classification tasks. Flexer argues that compared to traditional clustering methods, such as vector quantization and multidimensional scaling, "SOM performs significantly worse in terms of data points misclassified especially with higher numbers of clusters in the data sets" (1996, p. 446). Furthermore, traditional multidimensional methods tend to preserve the distances between inputs much more effectively than SOMs, as each variable relies on a predefined distance in feature space. It is also difficult to explain the results intuitively of some applications of SOMs, nor is it possible to build a generative model for the data (Villmann, 1999).

Given the purposes of the text clustering and the nature of the data, I suggest that the SOM model is still appropriate in our case and that the SOM model is more capable than traditional techniques in organizing large, complex datasets. The SOM can accurately define the similarity between data points, with positive impacts on clustering performance and the identification of authors of disputed texts (tweets). Furthermore, clustering is based on many different variables and features, including letter-pair frequencies as well as lexical properties, which are difficult to manage with traditional cluster analysis methods.

## 4. Data

This study is based on real-world data derived from tweets written by different Twitter users. Twitter was chosen because it is the world's largest microblog service. The tweets in this study have a maximum of 140 characters as they were retrieved shortly before Twitter officially expanded its character count to 280 on November 8, 2017. The shortness of these tweets made them appropriate for the purposes of this study. Furthermore, Twitter has a serious harassment and abuse problems due to the anonymous nature of many users. It is hoped that the results of this study can help in the detection of users who use social media platforms, such as Twitter, for illegal purposes. One problem encountered, however, was accessing the relevant data. Different free corpora, including the Edinburgh Twitter Corpus or Quandl, are no longer available. Furthermore, Twitter does not allow tweets to be published or shared online because of issues surrounding user rights. Furthermore, Twitter no longer allows tweets to be used for academic purposes for free. Acquiring Twitter data was not an entirely a straight-forward

process.

One way to overcome the challenge and obtain Twitter data was to directly retrieve data from the public Twitter Application Programming Interface (API). A piece of software was used to access the Twitter platform and acquire Twitter datasets as the API provides different functions for researchers, including extracting or retrieving tweets from user timelines. One advantage of this function is that every retrieved tweet is linked to its account or user, which was useful for cross validation purposes. There are however two main disadvantages with the API. First, it does not give access to historical data. Second, only a small portion of Twitter is available through its popular API, or other application programming interfaces. As such, data were only extracted from live streams, which were considered to be sufficient for the purposes of the study.

In order to limit the scope of the search, the topic 'Removal of Confederate monuments' was selected as a search term. Data was extracted during August 2017. At that time, tweets were limited to 140 characters. Tweets containing the words 'confederate monuments' were extracted. The tweets were in English, Spanish, and other languages. Only tweets written in English were used for the purposes of the study. Finally, a corpus of 12,240 tweets from 87 Twitter accounts was developed. Tweets per user ranged from 122 to 146, which was considered sufficient to construct an accurate author profile.

## 5. Analysis

The contents of tweets (letter-pair frequencies and lexical frequencies) were mathematically represented so that the data could be subjected to analysis and processing. GitHub was first used to extract letter combinations from the selected tweets. All consecutive letters anywhere within a word were extracted. All letter frequencies within words, whatever their position, were then identified and extracted. For example, the sentence 'the cat sat on the mat' would be segmented as 'th', 'he', 'ca', and so on. In this way, a list of all the two-character sequences, xy, is compiled and their relative frequency is computed. The number of each of the letter pairs (e.g., 'th', 'he', 'ca', and 'at') is counted for each of the selected tweets in the corpus. In our case, a list of all possible letter-pair combinations xy was generated. This gave a set of vectors (all possible occurrences of xy) for each of the selected tweets in the corpus. Following this, all lexical types were extracted.

For computing text similarity and assigning the selected tweets to their authors, SOM methods were used. The implementation of the SOM was carried out over two stages: training and mapping. The training phase involves adjusting the weight of the features or variables. Training addresses one of the most significant problems with text clustering applications, namely the high dimensionality of the data, which can have negative impacts on clustering performance. With too many dimensions, relative distances between the rows (documents) become meaningless and results are unreliable (Skillicorn, 2012). This is sometimes referred to as the curse of dimensionality (Blann, 2015; Ferraty & Romain, 2011). With large numbers of attributes or features, the number of dimensions can become staggeringly high making the calculations extremely difficult.

In the case of this study, there are thousands of variables covering the letter combinations and the distinctive lexical features of each text. The number of features or independent variables thus exceeds the number of observations and consequently the size of the space or context becomes unmanageable. As a solution, this study used SOMs for dimensionality reduction. There are a number of dimensionality reduction techniques available, but for this study the SOM technique was chosen because it ensures a reduced dimensional description that remains representative of the original body of data. Where the dimensionality of the data is very high and the number of objects makes classification difficult, the SOM first selects inputs in a random way, computes winner neurons (the most distinctive nodes/features), updates them, and repeats the process for all input data (Kohonen, 2012). The SOM thus provides an orderly mapping of a high dimensional space into much lower dimensional spaces, leading to dimension reduction and feature extraction for better classification performance (Chen, Lee, Kotani, & Ohmi, 2010). In this study, the high dimensions of the data were reduced through a process that produces winning nodes. This process is completed while preserving the neighborhood relationships that exist within the input datasets. The retained variables are the most distinctive features. These are included in a master list of all the unique variables of the datasets. This master list included 132 letter combinations and 145 lexical types.

A problem that arose in this process was the variation in document length. The selected tweets in this study, as with any given corpus, vary in length. This variation, if not addressed, can have negative impacts on clustering performance and reliability. Logically, documents that are longer have a greater number of words. As a result, the values or frequencies for those words increase, and a short document that may have high relevance for a given term will not necessarily have that relevance reflected in its term frequencies. Longer documents have higher term frequency values and naturally they have more distinct terms. The length factor inflates the scores of

longer documents, which distorts their impact. As such, longer documents may be favored simply because they have more terms. This leads to proximity measurements being dominated by longer documents. This means that if the length of the document increases, the number of times a particular term occurs in the document also increases. Consequently, length becomes an increasingly important determinant for clustering and these long documents will be clustered together. The same holds true if the documents are short—the angles between the vectors become smaller and as a consequence short documents will be clustered together.

The corpus of this study includes hundreds of tweets of variable length. Some tweets are composed of just one or two words (roughly 8–10 characters) while others are composed of 25–30 words (roughly 125–140 characters). If variation in document length is not addressed, long documents would have been ranked above short ones. To address this problem, mean document length normalization was used. This is one of the simplest and most straightforward normalization methods and involves the transformation of the row vectors of the data matrix in relation to the average length of documents in the corpus using the function:

$$M_i = M_i\left(\frac{\mu}{length(C_i)}\right)$$

Where:

M*i* is the matrix row representing the frequency profile of any document collection C,

*Length* (C*i*) is the total number of letter bigrams in C*i*, and

μ is the mean number of bigrams across all documents in C:

$$\mu = \sum_{i=1..m} \frac{length(C_i)}{m}$$

The values of each row vector M*i* are multiplied by the ratio of the mean number of bigrams per document across the collection C to the number of bigrams in document c*i*. The longer the document, the numerically smaller the ratio is, and vice versa. This has the effect of decreasing the values in the vectors that represent long documents, and increasing them in vectors that represent short ones, relative to average document length.

Having dealt with the data dimensionality and document length problems, the selected features are now ready for the next stage. In the mapping stage, similarities or common features between datasets are calculated and measured. For the purposes of this study, similarities between datasets are calculated and measured using Euclidean distances. Euclidean distance is the most commonly used distance measure—it is the most natural and intuitive way of computing a distance between two points and it is defined as the straight line distance between two points. In mathematical terms, Euclidean distance is concerned with studying the relationships between distances and angles in a space. According to Euclid, A 1-dimensional, 2-dimnsional, or 3-dimensional can be described and defined by axes. For a 1-dimensional space, only a single numerical measure is required. The distance between two objects can be defined by length and graphically represented, as in Figure 2.


Figure 2. Axis for a 1-dimensional space

Likewise, a 2-dimensional space can be defined using two numerical measures. A school's playground, for instance, can be defined in terms of length and width. The two measurements can be represented in Euclidean geometry as a 2-dimensional space as in Figure 3.

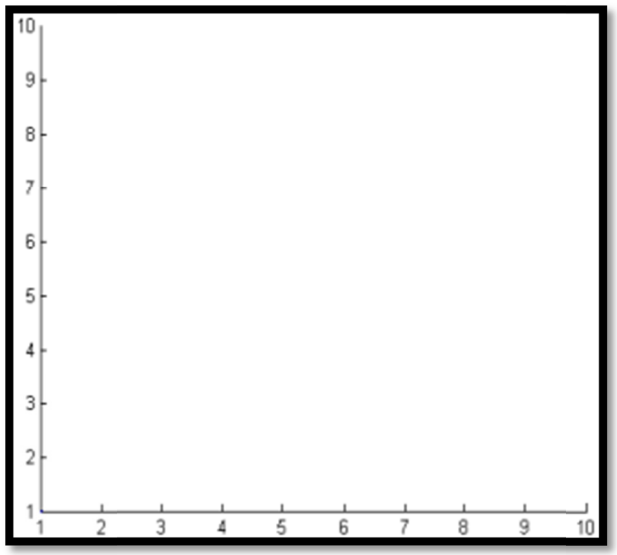


Figure 3. Axes for a 2-dimensional space

Euclid observed that there are still other kinds of physical property that cannot be described in one or two dimensions, but require three, such as real-world buildings. In such a case, three measurements are required: length, width, and height, and these can be represented in Euclidean geometry as a 3-dimensional space, as in Figure 4.

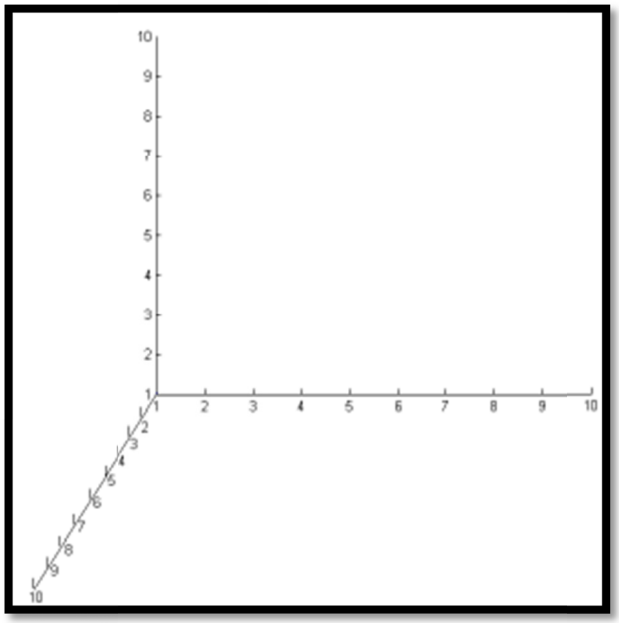


Figure 4. Axes for a 3-dimensional space

Because it was impossible to define more than three dimensions, modern mathematics generalized Euclid's concepts of distance, length, and angle so that any number of dimensions can be defined. For example, the economic growth of developing countries can be represented by an arbitrary large number of dimensions, such as the role of physical and human capital, technological progress, scale of investments, trade, capital mobility, fixed assets, net capital stock, and employment. These can be represented using N-dimensional space.

## 6. Results

As an initial step in assigning each document to its author, an SOM was used so that neurons (tweets) that shared the same morphological and lexical features were kept together in the same context or neighborhood. Feature maps or networks, where neurons in the same neighborhood have connections with each other and belong to a particular domain or feature, were constructed from this. The generated maps or networks were then used in classifying the tweets—the data, or tweets, were transformed through this mapping into a classification model. A neural network was developed for explaining how groups or classes were grouped geometrically. In these neural networks or maps, the best matching unit (BMU) between similar nodes was calculated using Euclidean distance methods and the nodes within the same neighborhood were determined. This process identified the clusters within the SOM by identifying each clusters component.

The SOM divided the space of the tweets into a number of clusters and every cluster or class included all those tweets that had a high coefficient of correlation. It is proposed that the further two clusters are from each other (the greater the difference in morphology and lexicon between two clusters), the lower the correlation coefficient between the tweets within these clusters is. The matrix divides into 8 main clusters and further divides into a number of sub-clusters. The number of these sub-clusters was compared to the same number of author profiles for validation. The mapped data points of each cluster were used in developing user segmentation profiles. The profile-based method was then used and all documents/tweets grouped together were considered to be written by the same author or user. Results obtained were then compared to the known-author tweets in order to find the correct authors of particular tweets and evaluate the performance of the proposed approach. The results indicated that classification accuracy based on the proposed system (using letter pair combinations and distinctive lexical features) is around 76%. Up to 22% of this accuracy was lost, however, when only distinctive words were used, and 26% was lost when the classification was based on letter combinations and morphological patterns only.

I suggest that the integration of letter-pairs and morphological patterns improves the accuracy of determining the authors of very short texts, as seen in the case of the Twitter posts analyzed here. This indicates that integrating the way the words are used along with the lexical features of the data, leads to better classification performance in analyzing very short texts. It is also clear that the use of a self-organizing map (SOM) leads to better clustering performance with its capacity to integrate two different linguistic levels (i.e., both morphological and lexical features) of each author profile. Unlike conventional classification methods, SOMs have the potential to integrate more than one variable with positive implications for identifying authorship. It was also clear that the data mapping was easily interpreted and that the tweets were clearly grouped and visualized in terms of the characteristics that unified them.

## 7. Conclusion

In order to address the limitations of current quantitative linguistic approaches to authorship detection in very short texts, this paper has proposed a new method that considers letter-pair frequencies/combinations along with the lexical features of documents. Given the uniqueness of language in social media, it is believed that letter information or mapping carries unique stylistic features that can be used alongside analysis of lexical features to enhance authorship detection in relation to very short texts. Controversial texts can thus be assigned to their authors by detecting stable word combinations and morphological patterns, as well as identifying lexical features. In order to test the proposed method, a corpus of 12,240 tweets derived from 87 Twitter accounts was created and the SOM model was used to classify input patterns that shared common features. This was used to assign tweets grouped under one class membership to a particular author. Results indicate that classification accuracy based on the integration of the morphological patterns and lexical features of texts is around 76%. Up to 22% of this accuracy was lost, however, when only distinctive words were used, and 26% was lost when classification was based on letter combinations and morphological patterns only. The integration of letter-pairs and morphological patterns had the advantage of improving the accuracy of determining the author of a given tweet. This indicates that the integration of different variables into an integrated system leads to a better classification performance when analyzing very short texts. It is also clear that the use of the self-organizing map (SOM) led to better clustering performance with its capacity to integrate two different linguistic levels (i.e., morphological and lexical features) of each author profile. It should be noted however that while this approach is suitable for tweets and very short texts (less than 140 characters) in English, it is not clear whether it is appropriate for other languages. It was also clear that the SOM model had the advantage of reducing the high dimensionality of data with minimal loss of information, which also had a positive impact on clustering performance. Finally, it can be claimed that the use of quantitative linguistics offers opportunities to detect linguistic properties and processes through computational methods. Such quantitative concepts can be usefully exploited to address the limitations of traditional linguistic and stylometric approaches. This study adopted a mathematical analysis of linguistic

properties and processes to assess authorship detection in relation to very short texts. This approach is consistent with the role that language technology and computational tools play in addressing scholarly issues regarding the changing nature of language in an era of social media.

**References**

Blann, A. (2015). *Data Handling and Analysis*. Oxford: Oxford University Press.

Brena, R. F. (2011). *Quantitative Semantics and Soft Computing Methods for the Web: Perspectives and Applications: Perspectives and Applications*. Information Science Reference. https://doi.org/10.4018/978-1-60960-881-1

Chaski, C. E. (2012). Author Identification in the Forensic Setting. In L. M. Solan & P. M. Tiersma (Eds.), *The Oxford Handbook of Language and Law*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199572120.013.0036

Chen, Q., Lee, F., Kotani, K., & Ohmi, T. (2010). *Face Recognition Using Self-Organizing Maps*. INTECH Open Access Publisher. https://doi.org/10.5772/9173

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). *Detecting Offensive Language in Social Media to Protect Adolescent Online Safety*. Paper presented at the International Conference on Privacy, Security, Risk and Trust, Sept. 3–5, 2012. https://doi.org/10.5772/9173

Coulthard, M., & Johnson, A. (2010). *An Introduction to Forensic Linguistics: Language in Evidence*. London and New York: Routledge.

Coulthard, M., & Johnson, A. (2013). *The Routledge Handbook of Forensic Linguistics*. London and New York: Routledge.

Craig, H. (2004). Stylistic Analysis and Authorship Studies. In S. Schreibman, R. Siemens & J. Unsworth (Eds.), *A Companion to Digital Humanities*. Oxford: Blackwell.

Davies, P., Francis, P., & Jupp, V. (2016). *Invisible crimes: their victims and their regulation*. Basingstoke: Macmillan Press.

Ferraty, F., & Romain, Y. (2011). *The Oxford Handbook of Functional Data Analysis*. Oxford: Oxford University Press.

Flexer, A. (1996). *Limitations of self-organizing maps for vector quantization and multidimensional scaling* (pp. 445–451). Advances in neural information processing systems, December 9.

Holland, J. (2017). Confederate statue toppled by protesters; more to be removed by cities. *The Mercury News*, August 16.

Johnsson, M. (2012). *Applications of Self-Organizing Maps*. InTech. https://doi.org/10.5772/3464

Juntunen, P., Liukkonen, M., Lehtola, M., & Hiltunen, Y. (2013). Cluster analysis by self-organizing maps: An application to the modelling of water quality in a treatment process. *Applied Soft Computing*, *13*(7), 3191–3196. https://doi.org/10.1016/j.asoc.2013.01.027

Kenning, C. (2017, August 28, 2017). Confederate Monuments Are Coming Down Across the United States. *The New York Times*.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*, 59–69. https://doi.org/10.1007/BF00337288

Kohonen, T. (1990). The Self-Organizing Map. *Proceeding of the IEEE*, *78*, 1464–1480. https://doi.org/10.1109/5.58325

Kohonen, T. (1995). *Self-Organizing Maps*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-97610-0

Kohonen, T. (2012). *Self-Organizing Maps* (3rd ed.). Berlin, Heidelberg: Springer.

Landrieu, M. (2018). *In the Shadow of Statues: A White Southerner Confronts History*. Penguin Publishing Group.

Liu, Y.-C., Liu, M., & Wang, X.-L. (2012). Application of Self-Organizing Maps in Text Clustering: A Review. In M. Johnsson (Ed.), *Applications of Self-Organizing Maps* (pp. 205–220). InTech. https://doi.org/10.5772/50618

Makagonov, P., Espinoza, C., & Sidorov, G. (2011). Document Search Images in Text Collections for Restricted

Domains on Websites. In R. F. Brena (Ed.), *Quantitative Semantics and Soft Computing Methods for the Web: Perspectives and Applications: Perspectives and Applications* (pp. 183–204). IGI Global. https://doi.org/10.4018/978-1-60960-881-1.ch009

Moisl, H. (2009). Using electronic corpora in historical dialectology research. In M. Dossena & R. Lass (Eds.), *Studies in English and European Historical Dialectology* (pp. 68–90). Brussels; Frankfurt: Peter Lang.

Nossel, S. (2017). The Problem with Making Hate Speech Illegal. *The Foreign Policy*, August 14.

Olsson, J. (2008). *Forensic Linguistics: An Introduction to Language, Crime and the Law*. London: Bloomsbury Publishing.

Olsson, J. (2009). *Word Crime: Solving Crime Through Forensic Linguistics*. London and New York: Continuum International Publishing Group.

Ostrowski, D. (2014). *Feature Selection for Twitter Classification* (pp. 267–272). IEEE International Conference on Semantic Computing, 16–18 June 2014. https://doi.org/10.1109/ICSC.2014.50

Savage, K. (2017). *Standing Soldiers, Kneeling Slaves: Race, War, and Monument in Nineteenth-Century*. America: Princeton University Press. https://doi.org/10.2307/j.ctt1tg5p86

Schreibman, S., Siemens, R., & Unsworth, J. (2004). *A Companion to Digital Humanities*. Oxford: Blackwell. https://doi.org/10.1111/b.9781405103213.2004.00003.x

Skillicorn, D. B. (2012). *Understanding High-Dimensional Spaces*. New York; London: Springer Science & Business Media. https://doi.org/10.1007/978-3-642-33398-9

Solan, L. M., & Tiersma, P. M. (2012). *The Oxford Handbook of Language and Law*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199572120.001.0001

Stolberg, S. G., & Rosenthal, B. (2017). Man Charged After White Nationalist Rally in Charlottesville Ends in Deadly Violence. *The New York Times*, August 12.

Sutton, M., & Mann, D. (1998). Net Crime: More Change in the Organisation of Thieving. *British Journal of Criminology*, *38*(2), 210–229. https://doi.org/10.1093/oxfordjournals.bjc.a014232

Timberg, C., & Harwell, D. (2018). We studied thousands of anonymous posts about the Parkland attack—and found a conspiracy in the making. *The Washington Post*, February 27.

Villmann, T. (1999). *Benefits and limits of self-organizing map and its variants in the area of satellite remote sensoring processing*. Paper presented at the ESANN'1999 proceedings - European Symposium on Artificial Neural Networks, Bruges (Belgium), April 21–23, 1999.

Wall, D. (2003). *Crime and the internet*. London: Routledge. https://doi.org/10.4324/9780203299180