Cybercrime and Authorship Detection in Very Short Texts

A Quantitative Morpho-lexical Approach

Abdulfattah Omar [1,2]

[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

Correspondence: Abdulfattah Omar, Department of English, College of Science & Humanities, Prince Sattam Bin Abdulaziz University, Al-Kharj, Riyadh, 11942, Kingdom of Saudi Arabia.

Abstract

The present study proposes an integrated framework that considers letter-pair frequencies/combinations along with the lexical features of documents. Drawing on a quantitative morpho-lexical approach, the study

tests the hypothesis that letter information or mapping carries unique stylistic features; and therefore detecting stable word combinations and morphological patterns can be used to enhance the authorship performance in relation to very short texts. The data used for analysis is a corpus of 12240 tweets derived from 87 Twitter accounts. Self-organizing maps (SOMs) model is used for classifying the input patterns that share common features together as a clue that tweets grouped under one class membership are written by the same author. Results indicate that the classification accuracy based on the proposed system 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 performance 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 linguistic variables into an integrated system leads to a better classification performance of very short texts. It is also clear that the use of the self-organizing map (SOM) led to better clustering performance for its capacity to integrate two different linguistic levels of each author profile together.

Key words: authorship detection, forensic linguistics, morphological patterns, lexical features, letter-pair frequencies, self-organizing maps (SOMs).

1. Introduction

With the world-wide development of computer and internet services throughout modern life, unprecedented problems and crimes have come to the surface with practically negative repercussions. These problems and crimes have already become the dark side of the Internet (Davies, Francis, & Jupp, 2016; Sutton & Mann, 1998; Wall, 2003). In social media applications, for instance, uncustomary illegal acts are committed in a way that poses real threats to the safe use of these applications. As a

corollary, different authorship identification techniques have been developed with the purpose of detecting the real identities of cyber criminals. However, different challenges in relation to the practical applications of authorship detection effectively addressed yet. These challenges and applications include identifying the authors of very short texts especially in social media applications. Practically, applications that tended to use conventional classification methods based on the lexical and/or structural properties did not usually yield reliable results in relation to the authorship detection of very short texts. Forensic text types are usually very short and thus have very minimal linguistic features. It may therefore be difficult for forensic linguists to adduce robust evidence due to the lack of sufficient linguistic data.

Addressing the foregoing problem, this study suggests a quantitative morpho-lexical approach that considers the two main variables of letter-pair frequencies as well as distinctive words and phrases for better authorship detection performance. The hypothesis is that authors usually have habits that are reflected unconsciously in their use of letters. Therefore, the analysis of author's style can readily be carried out through detecting stable word combinations in a given corpus (Brena, 2011; Makagonov, Espinoza, & Sidorov, 2011). The study of letter pair frequencies can thus be useful for the recognition of real authors of disputed texts. In other words, individuals have distinctive ways of writing as reflected essentially in the use of letters. It becomes thus a code or fingerprint by which authors can be revealed. The decoding of author's secret way of writing can thus lead to the identification of real authors of disputed texts. The problem with this approach, however, is that there are different variables which will be difficult for conventional cluster analysis to process. One way of solving this problem is the use of the self-organizing maps (SOMs) model due to its effectiveness in processing different variables simultaneously.

The current study is based on a corpus of selected tweets on the removal of the Confederate monuments in the United States in August 2017. In the United States, there are over seven hundred monuments across the country dedicated to the Confederate soldiers and leaders of the American Civil War who revolted against the US government's abolition of slavery. In 2015, some local governments in the United States made decisions

concerning the removal of these monuments as they were believed to represent white supremacy and racism. In August 2017, however, a white nationalist rally in Virginia renewed attention to the hundreds of the Confederate monuments around the country (Holland, 2017; Kenning, 2017). Supporters of the Confederate symbols were not happy with the planned removal of the Confederate monuments. They considered these monuments parts of the US history of which the great majority of American should be proud (Landrieu, 2018; Savage, 2017).

Inspired by the violent riots of the nationalists and conservatives in Virginia, counter-protesters, on the other hand, demanded the immediate removal of the confederate monuments and statues as symbols of racism and oppression. Some of them even did not wait for local officials to act and toppled a Confederate monument by themselves in Durham and several American cities. The political battles and controversial debates over the issue brought a quick flood of reaction on social media platforms including Twitter especially after the US President Donald Trump commented on the events on Twitter. He posted three tweets wherein he defended the Confederate monuments and described 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 the racism and hatred discourse among social media users (Nossel, 2017; Stolberg & Rosenthal, 2017). Furthermore, many observers linked the online hate speech to some real-life incidents. The topic thus provides an opportunity to extract real-life data for addressing one of the serious problems with Twitter and social media platforms. In our case, Twitter is used as an experimental case for testing a new authorship detection model based on the integration of letter-pair combinations as well as morphological and lexical features. As such, it can be assumed that the present study offers a proper site for investigating whether the authorship of very short texts can be detected using only linguistic stylometry.

2. Authorship detection and quantitative linguistics

The recent years have witnessed increasing rates of crimes associated with the use of social media networks. These included offensive language, hate messages, and even spreading terrorism and violence. It is true to say that instead of being platforms for social interaction; different social media networks have become effective facets for many abusers to post and send mean or embarrassing things about others; criminals to encourage hate crimes (e. g. religious, racist, and sexual orientation); and terrorists to spread their propaganda and inspire different people from different countries to commit different terrorist acts around the world. One main reason behind the spread of crimes of the kind is the anonymous nature of social networking or what can be described as the general potential for anonymity. Different social media applications and websites including Facebook and Twitter enable cyberbullies to send anonymous destructive messages to others that can result in character assassination or even suicide. It has been even revealed that the unanimous nature of Facebook and Twitter has been used to influence users' choices, spending habits, and even political decisions (e.g. the debates over the Russian intervention in the 2016 US elections). Although companies are constantly developing ways to deter and remove abusive posts, the damage of such posts usually remains. In the heart of this pseudonymity, different social media channels provide rare opportunities for attackers and cybercriminals to abuse others (by posting and sending verbal abuse, threats, 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 troubles and shape public understanding in particular ways. Timberg and Harwell (2018), for instance, argued that following the Parkland high school shootings in Florida, thousands of anonymous posts about the attack tended 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 its negative implications on society integrity. The idea that different social media networks have become a

potent tool of abuse and deception has made it more imperative to address the anonymity on the internet and think about novel and more 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 the applications to very short contents are not consistent. This applies to both linguistic and non-linguistic approaches to the problem. This study however is limited to the investigation of only linguistic approaches. 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 investigations 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, linguistic analysis has become more increasingly recognized by both researchers and investigation bodies since the second half of the twentieth century where forensic linguistics was initially developed (Chaski, 2012; Solan & Tiersma, 2012). The 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 is generally described as the application of linguistic knowledge, methods, and systems to legal settings. It tends to provide a careful and systematic analysis of language that can be used by different professionals including lawyers, judges, and jury members in evaluating questions of guilt and innocence in ways that serve justice and help find truth about crimes (Solan & Tiersma, 2012). With the development of computational methods, forensic linguistic approaches have become more reliable and forensic linguistics is considered today “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 linguistic fingerprint, which is defined as the process of collecting linguistic data and features which 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 for investigating the linguistic level/s chosen by the researcher. The majority of these quantitative or statistical linguistic approaches, known in the literature as stylometric approaches, are mainly based on the statistical investigation of the lexical, syntactic, and/or structural features of social media contents which has proved unsuccessful in detecting possible authors of offensive content. This is attributed to the idea that the language of the online social media is usually "highly unstructured, informal, and often misspelled" (Chen, Zhou, Zhu, & Xu, 2012, p. 71). Similarly, Ostrowski (2014) argues that the peculiar nature of social media language, being unorganized and characterized by the 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 the conventional stylometric approaches is that words are represented in the form of single words or n-grams (known in the literature as bag of words) using vector space model for measuring similarity between the 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 the texts, the lexical frequencies will be far too low and cluster analysis will generate spurious results. This leads to sparsity problems which have negative implications on results based on the frequency of lexical types. Authorship detection performance based on single words only is thus unreliable. The claim is that with the anonymous nature of internet applications today 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 the light of the limitations of the lexical and structural representations of text, this study suggests an integrated quantitative morpho-lexical approach that considers the use of letter pair frequencies along with the distinctive lexical features of texts for building a hierarchical cluster analysis with the purpose of successfully grouping similar texts together that can help in authorship identification performance. In traditional applications, documents were represented using single words only. The

rationale is 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, will be more easily identified. In this way, it is supposed that the use of the letter pair frequencies is appropriate for the nature of the data (the very sort social media contents). The research question, therefore, is asked in relation to the effectiveness of the use of letter pair frequencies in supporting the clustering performance and improving the authorship identification of anonymous users of social media networks.

3. Methodology

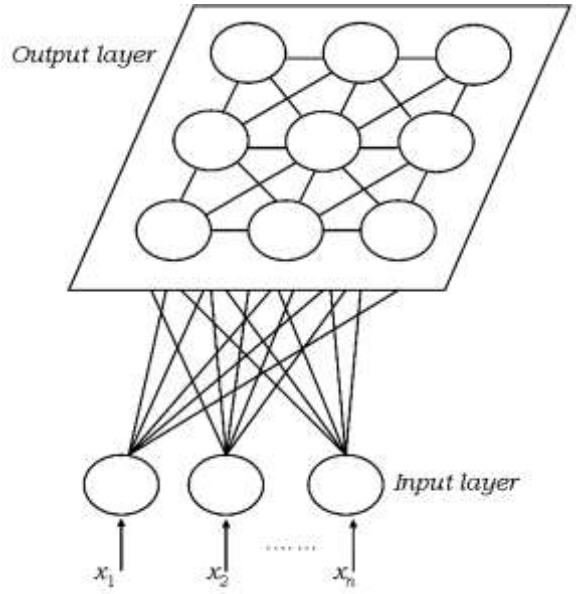
In order to address the problem of the authorship detection of very short texts, this study adopts a quantitative morpho-lexical approach. This is an integrated framework that considers both the quantitative morphological and lexical properties of texts. Although different authorship recognition systems 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 the stylometric information are thought to be effective clues for 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 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 used to be adopted at a very small scale till the closing years of the 20th century. Over the past two decades, however, quantitative linguistic methods have been widely used in order to address different problems related to natural language processing, machine translation, human intelligence, text classification, authorship detection, and information retrieval applications. Morphological and lexical analyses have been crucial in such applications.

The proposed technique is carried out at two subsequent stages. In the first stage, both the morphological and lexical information are extracted from the datasets (i.e. the tweets in our case) and graphically represented.

In other words, quantitative methods are used in order to capture only and all the distinctive morphological and lexical properties of the corpus and use them as inputs with the purpose of constructing a structural (graphical) representation that can be used in order to better understand the morphological patterns and the ways words are built. In our case, the hypothesis is that quantitative morpho-lexical methods are useful for identifying the distinctive morphological and lexical features and authors' writing style and thus assigning 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 in order to group similar texts together. The goal of ATC systems is to create clusters that are coherent internally, 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 maps (SOM) model is used. The model was first developed by Teuvo Kohonen in 1982 and it is now considered 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, reveal the similarity among the high dimensional data and map them onto a low dimensional map while keeping and retaining the distinctive features of the original datasets (Kohonen, 1982). IN SOM, the vectors, called here neurons or nodes, are arranged in a single, usually 2-dimensional grid. These represent the input layer. Neurons in the input layer then march out of the grid forming and through multiple iterations, successful neurons form areas with high density of data points which reflect 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 SOMs model has a number of advantages over other multivariate approaches including factor analysis and Principal Component Analysis (PCA). They explain that the SOMs 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 assumption is that the SOMs model is effective in enhancing the clustering performance and results as it makes advantages of the lexical properties as well as 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 the high dimensionality data space as well as keeping the most distinctive features which had their positive impacts on the clustering performance and accuracy, different problems and limitations have been raised when adopted it in some classification tasks. According to Flexer (1996, p. 446), 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". Furthermore, traditional multidimensional methods tend to preserve the distances much more effectively than SOM as the letter relies on a predefined distance in feature space. SOM applications also revealed that it is difficult to explain

the results intuitively and that it is not possible to build a generative model for the data (Villmann, 1999).

Given the purposes of the text clustering and the nature of the data, however, it is suggested that the SOM model is still appropriate in our case. The SOM model is more capable than traditional techniques for organizing large, complex datasets. To put it into effect, the SOM can accurately define the similarity between data points which will have its positive implications on the clustering performance and identifying authors of unanimous 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 be managed using traditional cluster analysis methods.

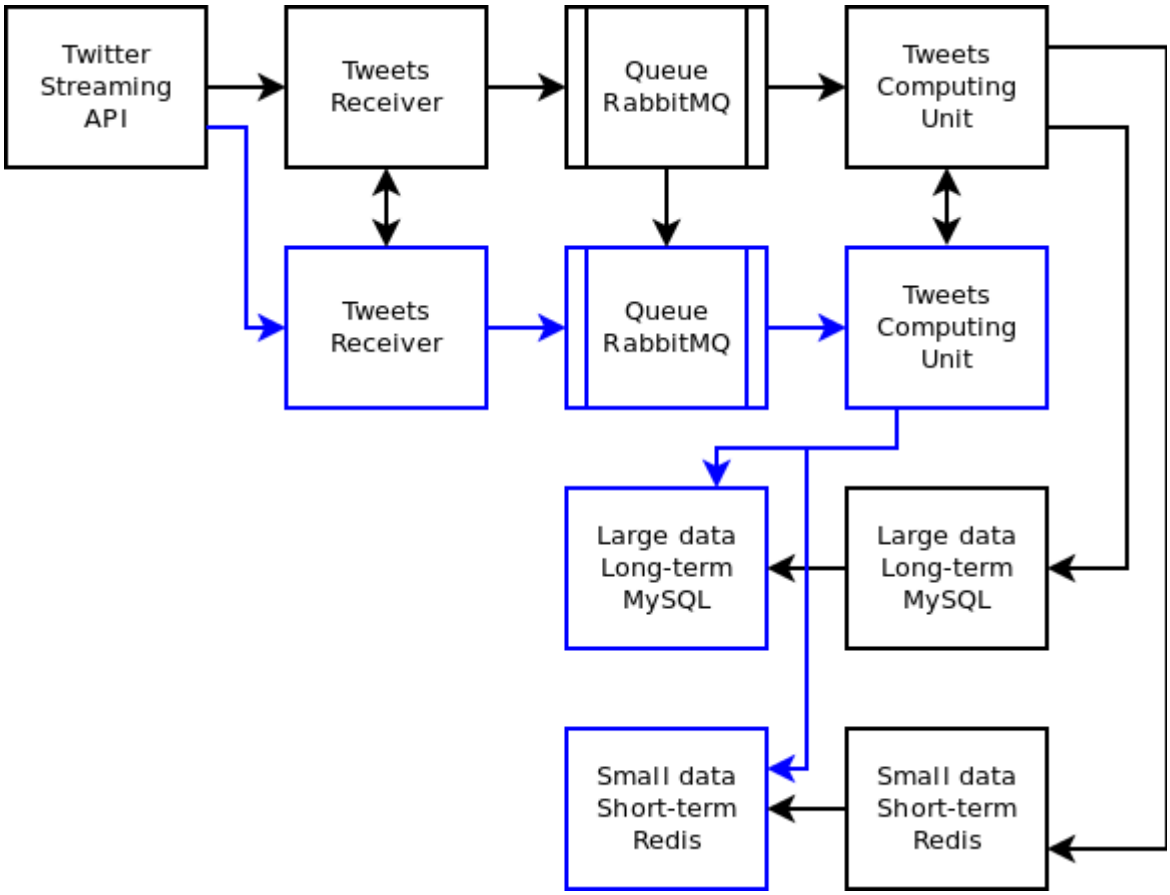
4. Data

For reliable results, the study is based on real-world data derived from tweets written by different Twitter users. The rationale is that Twitter is the largest microblog service. Furthermore, tweets are only 140 characters as a maximum since these were retrieved shortly before Twitter officially expanded its character count to 280 on November 8, 2017. The assumption is that tweets are very short so that they are appropriate for the purposes of the study. Furthermore, Twitter has serious harassment and abuse problems due to the unanimous nature of many users. It was thought then that the results of the study can help with the detection of users who use social media platforms and Twitter for illegal purposes. One problem, however, was accessibility to 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 for users' rights issues. Besides it no longer allows tweets to be used for academic purposes for free. Acquiring Twitter data thus is not entirely a straight-forward process.

One way to overcome the challenge and obtain Twitter data was to directly retrieve data from the Twitter public Application Programming Interface (API). In this way, a software was used to access the Twitter platform and acquire Twitter datasets. The rationale is that 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. This will be

useful for cross validation purposes. There are however two main disadvantages with the API. First it does not give access to historical data. In other words, it does not give access to collect data from the beginning of Twitter times. Second, only a small portion of Twitter is available through its popular API or any other application programming interference. Therefore, the data extraction was based only on live streams which were also thought to be really sufficient for the purposes of the study.

Figure 2: The architecture of Application Programming Interface (API)

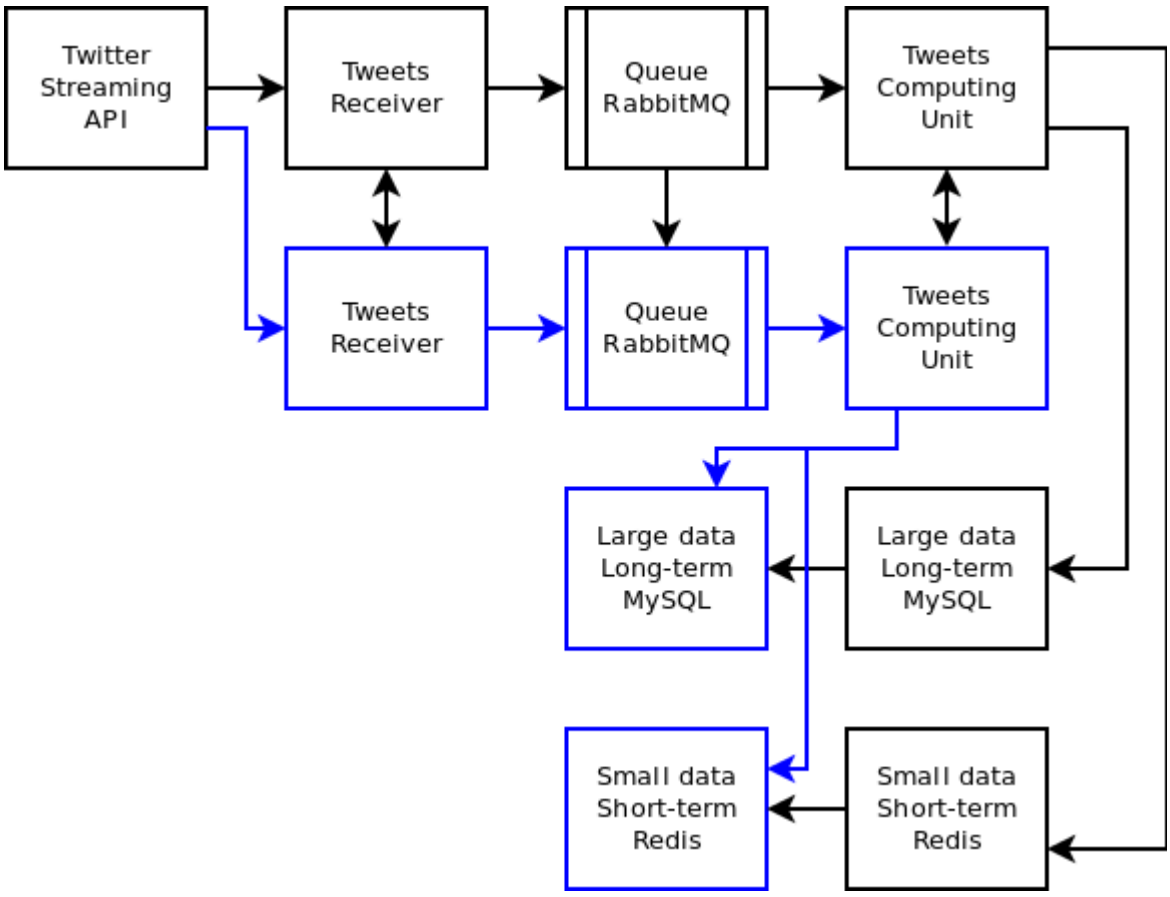


Source: (Laurent Luce, 2012)

In order to limit the search scope, the topic ‘Removal of Confederate monuments’ was selected. Data was extracted during August 2017. By that time, tweets were limited to just 140 characters. Tweets containing the two words ‘confederate monuments’ were extracted. The tweets consisted of English, Spanish, and different languages. Only tweets written in English were used for the purposes of the study. Finally, a corpus of 12240 tweets from 87 Twitter accounts was used for the

purposes of the study. Tweets per user ranged from 122 to 146 tweets which are considered appropriate for an accurate profile of an author.

Figure 3: Limiting tweets



Source: (Laurent Luce, 2012)

5. Application

The contents of tweets (letter pair frequencies as well as lexical frequencies) were mathematically represented so that data are amenable for analysis and processing. Gist Hub was first used for extracting the letter combinations of the selected tweets. All consecutive letters anywhere within a word were first extracted. All letter frequencies within words, wherever their positions are, were identified and extracted. Given the sentence 'the cat sat on the mat', it will 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. To each of the selected tweets in the corpus, the number of each of the letter pairs (e. g. ‘th’, ‘he’, ‘ca’, and ‘at’) is counted. In our case, a list of all possible letter-pair combinations *xy* was generated. The result is 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 SOMs was carried over two subsequent stages: training and mapping. The training phase is essentially based on adjusting the weight of the features or variables. In other words, it tends to address one of the most associated problems with text clustering applications, namely the high dimensionality of data. If not properly addressed, it usually has negative implications on the clustering performance. With too high dimensions, relative distances between the rows (documents) become meaningless and results are unreliable (Skillicorn, 2012). It is therefore some refer to high dimensionality of data as the curse of dimensionality (Blann, 2015; Ferraty & Romain, 2011). To put it simply, with large numbers of attributes or features, dimensions are staggeringly high so that calculations become extremely difficult.

In the case of the present study, there are thousands of variables. These are the letter combinations and the distinctive lexical features of each text. The number of features or the 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 self-organizing maps (SOMs) for dimensionality reduction. Despite numerous dimensionality reduction techniques are available, this study selects the SOM technique because it results finally in a reduced dimensional description which is representative of the original body of data Having a number of objects that are difficult to classify due to high dimensionality of data, 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 an input high dimensional space in much lower dimensional spaces, so it can play the role of dimension reduction and feature extraction for better classification performance (Q. Chen, Lee, Kotani, & Ohmi, 2010). In the case of the present study, the high dimensions of the data were reduced through a process known as the winning nodes. This process is done while keeping or preserving the neighborhood relationships that exist within the input datasets. The

retained variables are supposed to be the most distinctive features. These are included in what can be described as a master list of all and only the unique variables of the datasets. These included 132 letter combinations and 145 lexical types.

One more problem that came into the surface was the variation in document length. The selected tweets in the present study, like documents in any given corpus, vary in length. This variation, if not addressed, can have negative implications on clustering performance and reliability. Logically, documents that are longer have a higher number of words, hence the values or frequencies for those words are increased, and a document highly relevant for a given term that happens to be short will not necessarily have that relevance reflected in its term frequencies. Longer documents have higher term frequency values and naturally they have—for length reasons more distinct terms. The length factor results in raising the scores of longer documents, which is unnatural. So under the scoring scheme, longer documents are 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 of clustering and these long documents will be clustered together. Vice versa, if the documents are short, the angles between the vectors become smaller and as a sequence short documents will be clustered together.

The corpus of this study includes hundreds of tweets with variable length. Some tweets are composed of just one or two words (roughly 8-10 characters). Others are composed of 25-30 words (roughly 125-140 characters). If variation in document length is not addressed, long documents will be ranked above short ones. To address the problem, mean document length normalization is used. This is one of the simplest and most straightforward normalization methods. It involves 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

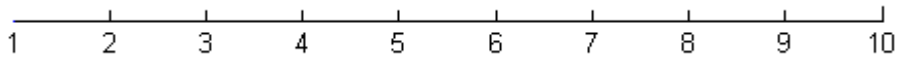
μ is the mean number of bigrams across all documents in C:

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

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.

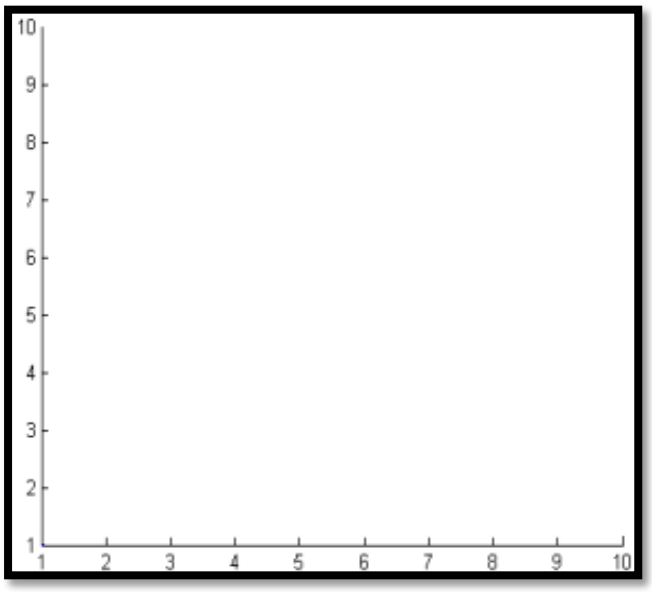
Having done with the data dimensionality and document length problems, the selected features are now ready for the next stage. In the mapping stage, similarities or common features between datasets are calculated and measured. For the purposes of the study, similarities between datasets are calculated 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. It is defined as the straight line distance between two points. In mathematical terms, Euclidean distance is concerned with studying the relationships among 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 4.

Figure 4: Axis for a 1-dimensional space



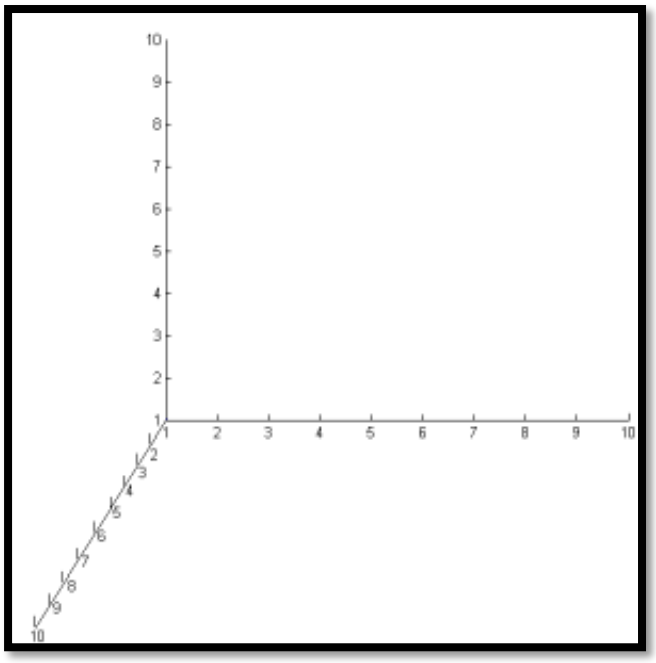
Likewise, a 2-dimensional space can be defined using two numerical measures. A school's playground, for instance, can be defined in terms of length and width. The two measurements can be represented in Euclidean geometry as a 2-dimensional space as in Figure 5.

Figure 5:Axes for a 2-dimensional space



Euclid observed that there are still other kinds of physical property which cannot be described in one or two dimension but require three, such as Big Ben Tower. 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 6.

Figure 6: 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. 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 for assigning each document to its author, SOM was used so that neurons (tweets) that share the same morphological and lexical features were kept close together in the same context or neighborhood. That resulted in building feature maps or networks where neurons in the same neighborhood have connections with each other and belong to a particular domain or feature. The generated maps or networks were then used for classifying the tweets. In other words, the datasets or tweets were transformed through this mapping into a classification model. A Neural Network was thus developed for explaining how groups or classes are grouped geometrically. In these neural networks or maps, Best Matching Unit (BMU) between similar nodes was calculated using Euclidean distance methods and the nodes within the same neighborhood

were determined. This process can also be described as identifying the clusters within the SOM where the components of each cluster were identified.

The SOM divided all the space of the tweets thus into a number of clusters and every cluster or class included all the tweets with high coefficient of correlation. The hypothesis is that the further two clusters from each other are (correspondingly the more difference in morphology and lexicon between two clusters is), the less the correlation coefficient between the tweets within these clusters are. The matrix falls into 8 main clusters which by turn fall into a number of sub-clusters. The number of these sub-clusters was decided to the same number of author profiles for comparison and validity purposes. Mapped data points of each cluster were thus used for developing user segmentation profiles. The profile-based method was then used where all documents/tweets grouped together were thought to be written by the same author or user. Results obtained were then compared to the known-author tweets in order to find correct authors of tweets and evaluate the performance of the proposed approach. Results indicate that the classification accuracy based on the proposed system (using letter pair combinations as well as 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 performance was based on letter combinations and morphological patterns only.

It can be claimed then that the integration of letter-pairs and morphological patterns had the advantage of improving the accuracy of determining the authors of very short texts as seen in the case of the Twitter posts. This indicates that the integration of the way words are built along with the lexical features of the data leads to a better classification performance of very short texts. It is also clear that the use of the self-organizing map (SOM) led to better clustering performance for its capacity to integrate two different linguistic levels (i.e. the morphological and lexical features) of each author profile together. Unlike conventional classification methods, SOM has the potentials of integrating more than one variable together which had ultimately its positive implications on the authorship performance. It was also clear that data mapping was easily interpreted. Tweets were clearly grouped and

visualized in terms of the uniformity of the characteristics that define them.

7. Conclusion

In order to address the limitations within the quantitative linguistic approaches to authorship detection of very short texts, this paper proposed a new method that considers letter-pair frequencies/combinations along with the lexical features of documents. Given the uniqueness of the social media language, it is believed that letter information or mapping carries unique stylistic features which can be usefully used along with the lexical features to enhance the authorship detection performance in relation to very short texts. Controversial texts can thus be assigned to their authors based on detecting stable word combinations and morphological patterns as well as identifying the most lexical features. In order to test the proposed method, a corpus of 12240 tweets derived from 87 Twitter accounts was created and the SOMs model was used for classifying the input patterns that share common features together. This was taken as a clue that tweets grouped under one class membership are written by the same author. Results indicate that the 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 the classification performance 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 of very short texts. It is also clear that the use of the self-organizing map (SOM) led to better clustering performance for its capacity to integrate two different linguistic levels (i.e. the morphological and lexical features) of each author profile together. 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 minimum loss of information which had its positive impact on the clustering performance. Finally, it can be claimed that quantitative linguistics has a good potential and offers opportunities

to detect linguistic properties and processes through quantitative and computational methods. It adopts such quantitative concepts that can be usefully exploited even to address limitations of traditional linguistic and stylometric approaches. Hence, the current study adopted the mathematical formulation of linguistic properties and processes to judge the quality of authorship detection and use them as tools to evaluate the performance of very short texts. This approach is consistent with the role that language technology and computational tools play to address the scholarly issues regarding the changing nature of language with the increased use of technology and 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.

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.

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

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. 3-5 Sept. 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 *Advances in neural information processing systems, 9*(December), 445-451.

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

Johnsson, M. (2012). *Applications of Self-Organizing Maps*: InTech.

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.

Kohonen, T. (1990). The Self-Organizing Map. *Proceeding of the IEEE, 78*, 1464-1480.

Kohonen, T. (1995). *Self-Organizing Maps*. Berlin, Heidelberg: Springer.

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.

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, 2017

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 *IEEE International Conference on Semantic Computing, 16-18 June 2014*, 267-272.

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.

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* Osford: Oxford University Press.

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

Sutton, M., & Mann, D. (1998). Net Crime: More Change in the Organisation of Thieving. *British Journal of Criminology, 38*(2), 210–229.

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, 2018

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) 21-23 April 1999

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