opción

Revista de Antropología, Ciencias de la Comunicación y de la Información, Filosofía, Lingüística y Semiótica, Problemas del Desarrollo, la Ciencia y la Tecnología

Año 34, 2018, Especial N° 18

Revista de Ciencias Humanas y Sociales
ISSN 1012-1587/ ISSNe: 2477-9385
Depósito Legal pp 198402ZU45

Universidad del Zulia
Facultad Experimental de Ciencias
Departamento de Ciencias Humanas
Maracaibo - Venezuela

Opción, Año 34, Especial No.18 (2018): 1765-1785
ISSN 1012-1587/ISSNe: 2477-9385

Cybercrime and authorship detection in very short texts

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

Aldawsari Bader Deraan[2]
[2]Department of English, College of Arts & Science, Prince Sattam Bin Abdulaziz University

Abstract

The aim of the study is to investigate cybercrime and authorship detection in very short texts via a quantitative morpho-lexical 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%. In conclusion, the use of the self-organizing map (SOM) led to better authorship performance for its capacity to integrate two different linguistic levels (i.e. the morphological and lexical features) of each author together, unlike other clustering systems.

Key words: Authorship Identification, Quantitative Morphology, Features.

Detección de ciberdelincuencia y autoría en textos muy cortos

Resumen

El objetivo del estudio es investigar la detección de la ciberdelincuencia y la autoría en textos muy cortos a través de un enfoque cuantitativo morfo-léxico. Los resultados indican que la precisión de la clasificación basada en el sistema propuesto (utilizando combinaciones de pares de letras y características léxicas distintivas) es de alrededor del 76%. En conclusión, el uso del mapa de

Recibido: 04-12--2017 •Aceptado: 10-03-2018

autoorganización (SOM) llevó a un mejor rendimiento de autoría por su capacidad para integrar dos niveles lingüísticos diferentes (es decir, las características morfológicas y léxicas) de cada autor, a diferencia de otros sistemas de agrupamiento.

Palabras clave: Identificación de autoría, morfología cuantitativa, características.

1. INTRODUCTION

With the development of computer and internet services which have changed the course of our modern life in many ways, some problems and a new kind of crimes have come into the surface which has their negative implications on the way we use these services. These problems and crimes represent what is described as the dark side of the internet. In social media applications, for instance, different illegal acts are committed in a way that poses real threats to the safe use of these applications. In response, different authorship identification techniques have been developed with the purpose of identifying the real identities of cybercriminals. However, different challenges in relation to the practical applications of authorship identification are not effectively addressed yet. These include identifying the authors of very short texts, especially in social media applications. To put it into effect, applications that tended to use conventional cluster analysis methods based on the lexical and/or structural properties did not usually yield reliable results in relation to the authorship recognition of very short texts.

In order to address the problem, this study suggests a quantitative morphological approach that considers two main variables: letter-pair

frequencies as well as distinctive words and phrases for better authorship identification performance. The hypothesis is that authors usually have habits that are reflected unconsciously in their use of letters. The analysis of author's style, therefore, can be carried out through detecting stable word combinations in a given corpus. The study of the frequencies of letter pairs 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. One problem with this approach, however, is that there are different variables which will be difficult for conventional cluster analysis to process. A solution is the use of the self-organizing maps (SOMs) model due to its effectiveness in processing different variables simultaneously.

By way of illustration, this 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 hundreds of 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 represent white supremacy and racism. It was thought these Confederate symbols represent white domination and superiority. 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 as parts of the US history which all American should be proud of. They argued that the removal of Confederate monuments is an indication of erasing and whitewashing history (Landrieu, 2018; Savage, 2017).

Inspired by the violent riots of the nationalists and conservatives in Virginia, outraged 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 (Astor, 2017).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 in which 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 …
>
> • … Cannot 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 between the online hate speech and 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.

2. LITERATURE REVIEW

The recent years have witnessed increasing rates of crimes associated with the use of social media networks. This 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 stylometry.

The majority of the quantitative 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 have proved unsuccessful in detecting possible authors of offensive content. This is attributed to the idea that the language of online social media is usually highly unstructured, informal, and often misspelled. 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 a bag of words) using vector space model for measuring the 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 morphological 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 is 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 (Sutton & Mann, 1998).

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 Kohonen in 1995 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. 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.

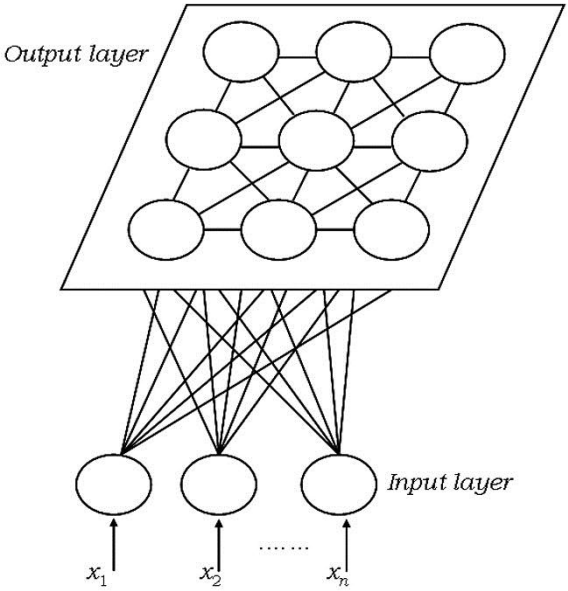


Figure 1: The way SOMs work

Juntunen et al. (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 takes advantages of the lexical properties as well as the relationship between letters (the letter combinations) in the input documents (Johnsson, 2012; Liu et al., 2012).

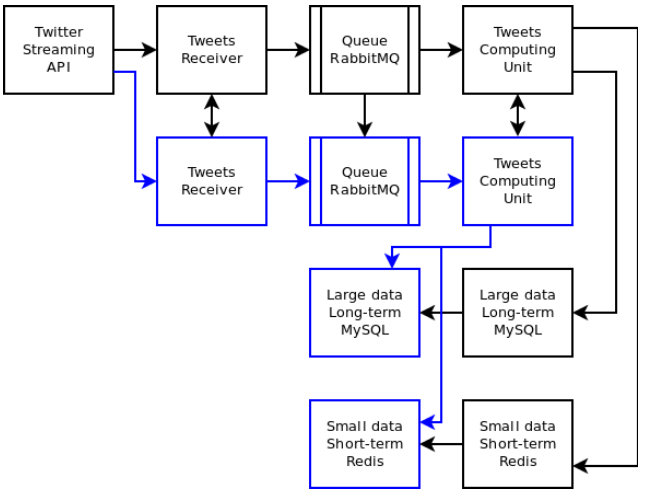
Given the purposes of the text clustering and the nature of the data, it is suggested that the SOM model is appropriate for the nature of the data and the purpose of data clustering. To put it into effect, the SOM can accurately define the similarity between data points which will have its positive implications 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 (Wall, 2003).

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 Quandlare 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. So acquiring Twitter data isnot 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, the 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. Two main disadvantages, however, with the API are that 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.

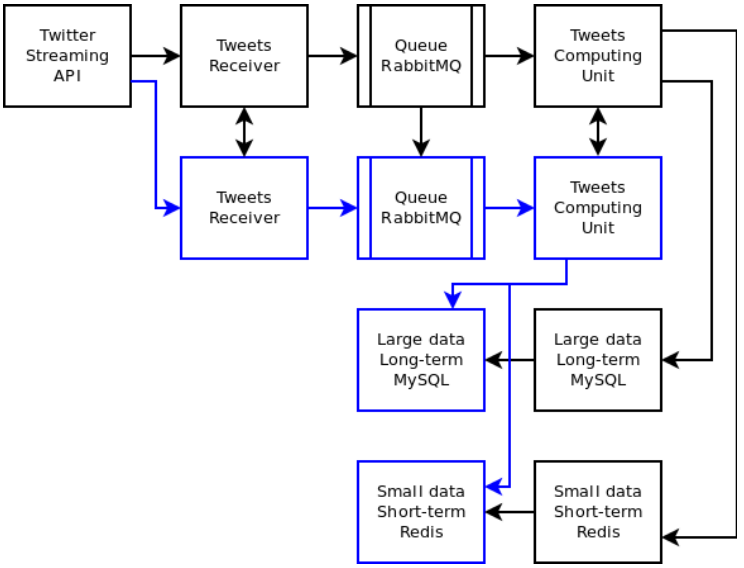
Figure 2: The architecture of Application Programming Interface (API)



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 87Twitter 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 has 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). That is why some refer to the high dimensionality of data as the curse of dimensionality. 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. 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 the 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 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 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.

6. RESULTS

In order to assign the documents to their plausible authors, the profile-based method was used where all documents/tweets for each author were grouped together. 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. 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 authorship performance for its capacity to integrate two different linguistic levels (i.e. the morphological and lexical features) of each author together, unlike other clustering systems.

7. CONCLUSION

This paper proposed a quantitative morpho-lexical approach for identifying the authors of very short texts. It was based on the assumption that controversial texts can be assigned to their authors based on both the vocabulary and the way this vocabulary is built. This new approach addressed the limitations of the previous quantitative linguistic approaches which were mainly based on the frequency of the occurrences of the words used in a given corpus. The morphological features and the ways words are arranged proved to be effective tools in improving the quality of authorship detection performance of very short texts. While this approach is suitable for tweets and very short texts (less than 140 characters) in English, it is not clear whether the proposed method is appropriate for other languages.

REFERENCES

ASTOR, M. 2017. **Protesters in Durham Topple a Confederate Monument**. The New York Times. USA.

HOLLAND, J. 2017. **Confederate statue toppled by protesters; more to be removed by cities**. The Mercury News. USA.

JOHNSSON, M. 2012. **Applications of Self-Organizing Maps**. InTech. UK.

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. Vol. 13, N^{o} 7: 3191-3196. Netherlands.

KENNING, C. 2017. **Confederate Monuments Are Coming Down Across the United States**. The New York Times. USA.

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

LANDRIEU, M. 2018. **In the Shadow of Statues: A White Southerner Confronts History**. Penguin Publishing Group. USA.

LIU, Y., LIU, M., & WANG, X. 2012. **Application of Self-Organizing Maps in Text Clustering: A Review**. In M. Johnsson (Ed.), Applications of Self-Organizing Maps. pp. 205-220. InTech. UK.

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. Belgium.

NOSSEL, S. 2017. **The Problem with Making Hate Speech Illegal**. The Foreign Policy. USA.

OSTROWSKI, D. 2014. **Feature Selection for Twitter Classification.** IEEE International Conference on Semantic Computing. pp. 267-272. USA.

SAVAGE, K. 2017. **Standing Soldiers, Kneeling Slaves: Race, War, and Monument in Nineteenth-Century America**. Princeton University Press. USA.

SKILLICORN, D. 2012. **Understanding High-Dimensional Spaces**. Springer Science & Business Media. Germany.

STOLBERG, S., & ROSENTHAL, B. 2017. **Man Charged After White Nationalist Rally in Charlottesville Ends in Deadly Violence**. The New York Times. USA.

SUTTON, M., & MANN, D. 1998. **Net Crime: More Change in the Organization of Thieving**. British Journal of Criminology. Vol. 38, Nº 2: 210–229. UK.

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. USA.

WALL, D. 2003. **Crime and the internet**. London: Routledge. UK.

UNIVERSIDAD DEL ZULIA

opción
Revista de Ciencias Humanas y Sociales

Año 34, Especial N° 18, 2018

Esta revista fue editada en formato digital por el personal de la Oficina de Publicaciones Científicas de la Facultad Experimental de Ciencias, Universidad del Zulia.
Maracaibo - Venezuela

www.luz.edu.ve

www.serbi.luz.edu.ve

produccioncientifica.luz.edu.ve