Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 175 (2020) 114–119

Procedia Computer Science

www.elsevier.com/locate/procedia

The 17th International Conference on Mobile Systems and Pervasive Computing (MobiSPC)
August 9-12, 2020, Leuven, Belgium

Analysis of behavior of automatic learning algorithms to identify criminal messages.

Noel Varela[a]*, Jesús Arturo Gálvez Valega[b], Omar Bonerge Pineda Lezama[c]

[a,b] *Universidad de la Costa, Barranquilla, Colombia*
[c] *Universidad Tecnológica Centroamericana (UNITEC), San Pedro Sula, Honduras*

Abstract

In this type of explanation, strictly economic or criminal motives predominate: mainly the control of routes and places, and the punishment of desertion or treason. The precarious and fragmentary nature of the public discourse of drug traffickers as well as the preponderance of police narratives has concealed the strictly political dimension of "criminal" violence in Colombia. In pragmatic terms, organized crime and politics are more similar than we would like to assume. They have in common the objective of dominating territories, resources and populations; both tend to stand as a system of "parasitic intermediation". Both mafias and the state offer "protection" in exchange for payment of fees, reward loyalty and punish treason. It is the discursive acts that accompany violence and the series of institutional procedures in which they are registered that allow us to draw the line between the political and the criminal, the legitimate and the illegitimate, the just and the unjust. In Colombia, that border has lost clarity. In this study, an analysis of narco-messages found in banners, social networks and other databases is carried out by applying data mining, in order to propose a geospatial model through which it is possible to identify and geographically distribute the authors of the messages.

Keywords: Text analysis model; Identification of authors; Criminal messages.

* Corresponding author. Tel.:+57-3235810446.
E-mail address: nvarela2@cuc.edu.co

10.1016/j.procs.2020.07.019

1. Introduction

In Colombia, municipal and state government agencies have used criminal groups to impose political control, and there have been reports of employees moving between the municipal police and private armed groups. In addition, in recent years there has been a growing participation of members or former members of organized crime in electoral politics. But there is another, perhaps more subtle, dimension to this approach that has to do with the difficulty of the state in establishing and defending what, in principle, distinguishes it from other armed groups. The difficulty of discursively drawing the line between crime and politics [1].

This loss of authority implies that the series of discursive acts that constitute the daily practice of the State from the granting of a driver's license to the resolution of a judicial investigation have been losing linguistic effectiveness [2][3]: capacity to affect the world. It implies that government agencies face more and more problems when trying to establish themselves as reliable sources. It is difficult to say that there are a series of specific political demands for drug trafficking, as was the case, for example, with the fight against extradition in Colombia. Nor does there seem to be a coherent and general social or ideological narrative that frames, defines or gives meaning to the suffering. There is, for example, no discourse that allows pain to become a sacrifice oriented towards the achievement of a greater good, since it does not guarantee the survival of the next generation: "I got involved in this so that my children do not have to kill their backs by working". It is not enough to form a political subject as such, an "us" well defined with its own demands as in [4].

Even so, in the sporadic and somehow unsuccessful public expressions of drug trafficking, it is possible to delineate something that goes beyond the strictly economic or criminal and that suggests the ideological and political dimensions of the conflict. This paper analyzes a particular type of expression: messages written on pieces of cloth or cardboard that began to appear frequently on the public streets around 2006. Narco-messages are almost more media than the message: their form goes beyond the meaning and content [5]. First, because many derived their public visibility and discursive force from appearing physically associated with a corpse. It is not only the context in which the blanket is found, but also its form. The vast majority are written in spray paint, with abundant spelling errors, insults, and unintelligible statements. The exception to this rule has been the blankets of criminal organizations.

2. Methodology

As a result of the need to contribute to improving security in Colombia, automated methods for analyzing message content and identifying potential authors are increasingly essential [6]. In this context, this research seeks to make a contribution to the national police (PN) in identifying potential perpetrators of these crimes by analyzing the content of messages using natural language processing and AI techniques.

The existing problem is largely due to the following characteristics [7]:

- Insufficient human resources.
- Large amount of information available (message characteristics)
- There is no articulated automation mechanism.

Human beings have unique habits and patterns of behavior, which leads to conjecture that certain characteristics (common words, spelling mistakes, autograph signatures, among others) will be constant. Therefore, the type of message and the focus of the contribution of its impact will be determined: categorization of the author. Based on the semantics of the message, it is possible to determine the purpose of the message (threatening, claiming territory, revenge, among others). Therefore, it is important to analyze, design and implement a mechanism for the NP, in order to support the identification of the geographical distribution of message authors using techniques such as natural language processing and text mining. To do so, it is necessary to obtain a set of characteristics from a series of messages for their analysis, sort a set of messages by means of similarities to assign them to a specific author, and then generate distribution maps where those authors operate. This way, the NP will have a model that will allow it to efficiently automate the analysis of the messages' content, as well as to identify the authors and group them geographically [8][9].

3. Text analysis and geo-location

For text analysis and geolocation, a three-phase model is proposed, which is described below [10][11][12].

Phase 1: Repository generation.

- A sample of 100 drug messages is obtained and analyzed with social data mining techniques.
- An image processing (Selection of images with readable text) is performed to determine the location of the images on a map of the city.
- OCR (optical character recognition) is applied with Matlab to convert a scanned text image into a text document.

Phase 2: Identification of the authorship using the Weka tool.

- In this phase, criminal groups are incorporated into the repository in order to extend the message corpus. Through similarity measures (e.g., distance from Mahalanobis) in the messages, the possible "criminal group" will be identified, generating agglomerations.

Phase 3: Generation of criminal group distribution maps to determine similar robbery-violence scenarios.

The data obtained and stored in the repository shall be processed using the R language to determine the frequency of items related to the same criminal group.

- The results will be discussed in the light of similar studies, in order to propose a public social policy to support people who require more protection.

4. Application of tools

The WEKA [13][14][15] social data mining tool was used to analyze the corpus data, based on the following process. First, a model was developed to explain the behavior of three samples of people, and how it affects their style of discourse related to narco-messages in narco-blankets. Among the results obtained with WEKA, a relationship was found between the parameters of hypostasis and parataxis, used by the different readers of this message, the speakers communicate with the Criminal group [16][17][18].

Table 1. Distribution of claims by category and ordered by three analyzed samples.

	Sample 1	**Sample 2**	**Sample 3**
Language	Influence	Challenges	Threats
N	521	254	412
Imperatives	14%	38%	25%
Directive statements	7%	8%	8%
Simulation directives	9%	5%	6%
Interrogative Directives	3%	1%	3%
Postscripts of interrogatives	33%	20%	30%
Joint Directive	14%	4%	15%
Explosive questions	4%	13%	1%
Information questions	18%	24%	11%
Mechanisms for attracting attention	**1%**	**1%**	**3%**

Also, it is found in both cases that users and readers of these messages showed a higher hypostasis and lower parataxis with respect to Spanish speakers. This can be explained by the use of informal speech of people related to "narcocorridoes", songs related to criminal groups, because they try to assimilate more easily to people with common ancestors (several people in this criminal group are native speakers of the same region in Colombia), and the decision to buy is highly influenced by the language community.

5. Text analysis and geo-location

It considers a sample of 587 message segments (521 Influence messages, 254 Challenge messages and 412 Threat messages) related to criminal groups recovered in the last three years, made up of three samples (sample 1 with Influence messages, sample 2 with Challenge messages and sample 3 with Threat messages), as well as conversations in social networks, to identify different behaviors (see Table 1).

The use of data mining on social issues has proven to be a key part of corroborating the linguistic trends of an established group within a common social network, however, it is possible to find certain variations depending on the intention of the message and the linguistic resource used in different languages, see Table 2.

Table 2. Contributions made to the discourse by a social network according to different words, including the turns of phrase used by the language.

	Volume of Speech		
Type of message	**Emitted words**	**Repetitions**	**Average of words in a time**
Influence	1254	521	6.3
Challenges	3251	365	8.5
	652	**254**	**6.1**

Finally, with the location of each drug message, a geospatial model is proposed to represent each scenario and determine future situations related to this type of criminal group. In Figure 1, this model is presented on a map of the department of Risaralda in Colombia. It shows the average number of words per region.

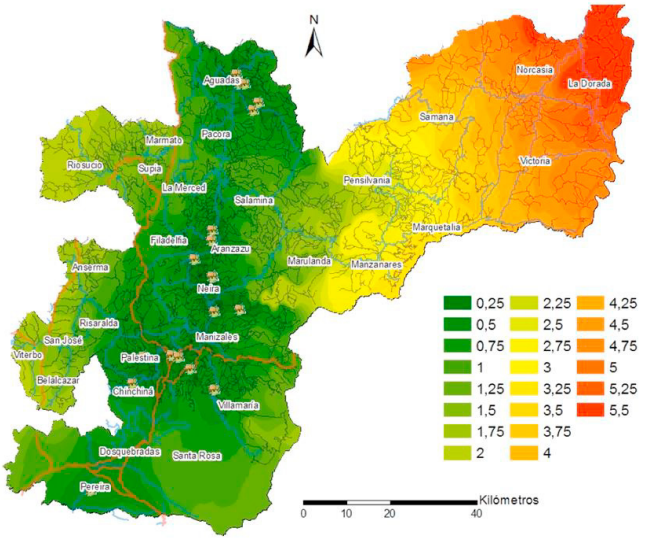

Fig. 1. Geospatial model with the locations of each narcomessage in Risaralda and the places where it is more specific that a new message is produced.

6. Conclusions

There are a number of important questions that deserve further investigation. One of them would be to find new sources of information on the use of these three languages and other cities with similar problems of criminal situations as in the city of Medellín (where in a range of 720 days it had just over 100 narcotics) [19]. One area with great potential is the use of electronic media, specifically digital music [20]. [21] shows a system that learns from user preferences based on the music played. After songs are selected to be played in a shared physical environment, based on the preferences of all the people present, this software has a narrative script to make recommendations to other users in a free text [22][23][24].

References

[1] Srinivasan, L., & Nalini, C. (2019). An improved framework for authorship identification in online messages. Cluster Computing, 22(5), 12101-12110.

[2] Manek, A. S., Shenoy, P. D., Mohan, M. C., & Venugopal, K. R. (2017). Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. World wide web, 20(2), 135-154.

[3] Gottlieb, A. (2017). The effect of message frames on public attitudes toward criminal justice reform for nonviolent offenses. Crime & Delinquency, 63(5), 636-656.

[4] Iqbal, F., Fung, B. C., Debbabi, M., Batool, R., & Marrington, A. (2019). Wordnet-based criminal networks mining for cybercrime investigation. IEEE Access, 7, 22740-22755.

[5] Duarte, N., Llanso, E., & Loup, A. (2018, January). Mixed Messages? The Limits of Automated Social Media Content Analysis. In FAT (p. 106).

[6] Kounadi, O., Ristea, A., Leitner, M., & Langford, C. (2018). Population at risk: using areal interpolation and Twitter messages to create population models for burglaries and robberies. Cartography and geographic information science, 45(3), 205-220.

[7] Ajao, O., Bhowmik, D., & Zargari, S. (2018, July). Fake news identification on twitter with hybrid cnn and rnn models. In Proceedings of the 9th International Conference on Social Media and Society (pp. 226-230).

[8] Barbon, S., Igawa, R. A., & Zarpelão, B. B. (2017). Authorship verification applied to detection of compromised accounts on online social networks. Multimedia Tools and Applications, 76(3), 3213-3233.

[9] Venckauskas, A., Karpavicius, A., Damaševičius, R., Marcinkevičius, R., Kapočiūte-Dzikiené, J., & Napoli, C. (2017, September). Open class authorship attribution of lithuanian internet comments using one-class classifier. In 2017 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 373-382). IEEE.

[10] Tundis, A., & Mühlhäuser, M. (2017, October). A multi-language approach towards the identification of suspicious users on social networks. In 2017 International Carnahan Conference on Security Technology (ICCST) (pp. 1-6). IEEE.

[11] Tundis, A., & Mühlhäuser, M. (2017, October). A multi-language approach towards the identification of suspicious users on social networks. In 2017 International Carnahan Conference on Security Technology (ICCST) (pp. 1-6). IEEE.

[12] Tracy, S. J. (2019). Qualitative research methods: Collecting evidence, crafting analysis, communicating impact. John Wiley & Sons.

[13] Soundarya, V., Kanimozhi, U., & Manjula, D. (2017). Recommendation System for Criminal Behavioral Analysis on Social Network using Genetic Weighted K-Means Clustering. JCP, 12(3), 212-220.

[14] Chang, V. (2018). A proposed social network analysis platform for big data analytics. Technological Forecasting and Social Change, 130, 57-68.

[15] Tundis, A., Bhatia, G., Jain, A., & Mühlhäuser, M. (2018, November). Supporting the identification and the assessment of suspicious users on twitter social media. In 2018 IEEE 17th International Symposium on Network Computing and Applications (NCA) (pp. 1-10). IEEE.

[16] Tseng, T. Y., Krebs, P., Schoenthaler, A., Wong, S., Sherman, S., Gonzalez, M., ... & Shelley, D. (2017). Combining text messaging and telephone counseling to increase varenicline adherence and smoking abstinence among cigarette smokers living with HIV: a randomized controlled study. AIDS and Behavior, 21(7), 1964-1974.

[17] Tundis, A., Jain, A., Bhatia, G., & Muhlhauser, M. (2019, July). Similarity Analysis of Criminals on Social Networks: An Example on Twitter. In 2019 28th International Conference on Computer Communication and Networks (ICCCN) (pp. 1-9). IEEE.

[18] DeHart, D., Dwyer, G., Seto, M. C., Moran, R., Letourneau, E., & Schwarz-Watts, D. (2017). Internet sexual solicitation of children: a proposed typology of offenders based on their chats, e-mails, and social network posts. Journal of sexual aggression, 23(1), 77-89.

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

[20] Sarna, G., & Bhatia, M. P. S. (2017). Content based approach to find the credibility of user in social networks: an application of cyberbullying. International Journal Of Machine Learning and Cybernetics, 8(2), 677-689.

[21] Scrivens, R., Davies, G., & Frank, R. (2018). Searching for signs of extremism on the web: an introduction to Sentiment-based Identification of Radical Authors. Behavioral sciences of terrorism and political aggression, 10(1), 39-59.

[22] Macnair, L., & Frank, R. (2018). The mediums and the messages: Exploring the language of Islamic State media through sentiment analysis. Critical Studies on Terrorism, 11(3), 438-457.

[23] Kamatkar, S. J., Tayade, A., Viloria, A., & Hernández-Chacín, A. (2018). Application of classification technique of data mining for employee management system. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 10943 LNCS, pp. 434–444). Springer Verlag. https://doi.org/10.1007/978-3-319-93803-5_41

[24] Balaguera, M. I., Vargas, M. C., Lis-Gutierrez, J. P., Viloria, A., & Malagón, L. E. (2018). Architecture of an object-oriented modeling framework for human occupation. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 10942 LNCS, pp. 452–460). Springer Verlag. https://doi.org/10.1007/978-3-319-93818-9_43