Enhancing the Detection of Criminal Organizations in Mexico using ML and NLP

Javier Osorio
School of Government and Public Policy
University of Arizona
Tucson, AZ, United States
josorio1@email.arizona.edu

Alejandro Beltran
School of Government and Public Policy
University of Arizona
Tucson, AZ, United States
alejandrobeltran@email.arizona.edu

***Abstract*—This paper relies on Machine Learning (ML) and supervised Natural Language Processing (NLP) to generate a geo-referenced database on the violent presence of Mexican Criminal Organizations (MCOs) between 2000-2018. This application responds to the need for high-quality data on criminal groups to inform academic and policy analysis in a context of intense violence such as Mexico. Powered by ML and NLP tools, this computational social science application processes a vast collection of news stories written in Spanish to track MCOs' violent presence. The unprecedented granularity of the data allows disaggregating daily-municipal information for 10 main MCOs comprising more than 200 specific criminal cells.**

***Index Terms*—text classification, machine learning, information extraction, event data, maps**

I. Introduction

Conflict scholars and policy makers in the security and law enforcement sectors require fine-grained and timely information to track and understand the dynamics of violence in highly complex and rapidly changing scenarios of intense conflict [1] [2]. Unfortunately, such need is often hampered by the slow pace, limited scope, and high cost of manually generated data [3]. This problem is even more acute in developing countries thorn by violence where the availability of resources for research an analysis is scarce [4] [5] [6].

By focusing on the escalation of crime related violence in Mexico, this paper advances new Machine Learning (ML) protocols and Natural Language Processing (NLP) tools in Spanish to generate a geo-referenced database on the violent presence of Mexican Criminal Organizations (MCOs) between 2000-2018. To generate this database, the methodological strategy takes advantage of a large collection of articles from local and national newspapers and press releases from a variety of government agencies. The resulting database presents geo-referenced data at unprecedented levels of granularity by comprising daily-municipal information of 10 main criminal organizations that can be further disaggregated into more than 200 criminal cells. The application also presents a new interactive web interface presenting dynamic heat-maps of criminal territories. This information is used to identify distinct temporal and spatial trends in MCO development and contestation during this time period.

This research was possible thanks to the generous support of the Technology and Research Initiative Fund of the University of Arizona, and earlier grants from the National Science Foundation [SES-1123572], the Harry Frank Guggenheim Foundation, and the Drug Security and Democracy Fellowship of the Social Science Research Council – Open Society Foundations.

The paper is divided into four parts. The next section briefly reviews advances in machine-generated data for conflict research and addresses recent efforts on tracking criminal violence in Mexico using computerized approaches. The following section explains the methodology using Machine Learning for news article classification and Natural Language Processing for generating geo-referenced data of criminal groups. The subsequent part presents temporal trends of each major MCO between 2000 and 2018 along with maps of their territorial distribution throughout Mexico. The final section concludes with a highlight on the relevance of this research on event data and violence in Mexico.

II. Literature review

A. Text as data in the Social Sciences

In recent years, developments in Computational Social Science facilitated a massive production of machine-coded event data for conflict research [3] [1] [7] [8] [2] [9]. These efforts opened the possibility of advancing research on the determinants, characteristics, and prospective of violent conflict as well as informing policy makers about emerging threats.

Of course, the computerized generation of event data for conflict research is not free from the limitations of information sources [10] [11], and geo-referencing the precise location of specific incidents still is a challenge [12] [13]. However, despite these difficulties, the generation of computer-based approaches have made important contributions to producing accurate and valid information about conflict processes.

Text-as-data is becoming a popular approach for empirical analysis in the social sciences. The number of available text sources is constantly increasing with the digitization of texts and the availability of web sources. For example, scholars use text-as-data to identify policy changes, actors, and topics relevant to voters [14]. In their summary of text-as-data in political science, Wilkerson and Casas [15] discuss the strengths and limitations of these tools applied to the social sciences. Analyzing text-as-data is also gaining popularity in the public administration literature as summarized by Hollibaugh [16].

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Researchers also frequently use news articles as a source for identifying the occurrence of events around the world. Geo-referenced event data is an important tool for social scientists interested in researching political and social processes, including the escalation and spread of political violence [17]–[19]. Popular event datasets for studying conflict based on text-as-data include the Armed Conflict Location and Event Dataset (ACLED) [20] and the Uppsala Conflict Data Program (UCDP) [21], as well as computerized event data generators such as ICEWS [22] and the Phoenix event data project [23].

B. Applications to Mexico

Text-as-data has also been used to study MCO's dynamics in Mexico. Osorio [24] developed the Organized Criminal Violence Event Data (OCVED) that uses Spanish language news articles collected from 2000 to 2010 to identify MCO activity at the municipal level. OCVED relies on manually collected news articles and relies on sparse-parsing using Eventus ID [25] to find matches of MCO's actions in the text. This article advances OCVED by updating its temporal coverage up to 2018 and implementing ML tools in the news gathering and classification process.

Coscia and Rios [26] reduce the burden of manually collecting news articles about MCO's by scraping news stories using a web crawler through Google News between 2000 and 2010. Sobrino [27] also uses a web crawler to collect information on MCO's from 1990 to 2016. Then, she uses a Convolutional Neural Network text classifier to identify sentences mentioning both a cartel and a municipality to reduce noise in the identification of locations collected through the crawler. This method yields a classification accuracy of 86%. Machine Learning is a novel approach that significantly decreases the time required to generate MCO data while ensuring high accuracy in the classification task. The next generation of MCO data generation should extend ML applications combined with named entity recognition, universal dependencies, and event extraction to identify trends in criminal violent behavior rather than exclusively focusing on cartel territoriality.

III. ML AND NLP PROCESSES

To track the violent presence of Mexican criminal groups across time and space, this paper performs four main tasks described in Fig. 1: (*A*) automated web scraping; (*B*) news story classification using ML; (*C*) event coding using NLP in Spanish; and (*D*) data visualization using time series (TS) and Geographic Information Systems (GIS).

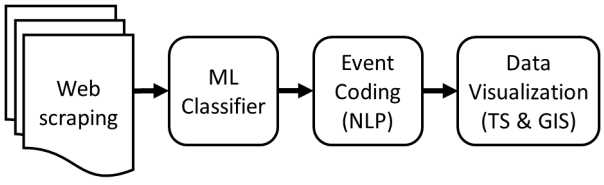


Fig. 1: ML and NLP process

A. Automated web scraping

The information used in this application comes from EMIS University, a global news aggregator of newspapers in multiple languages. Using EMIS' search engine, we run a query over 77 Mexican newspapers published in Spanish between 2010-2018 to identify MCO-related articles, returning thousands of news stories. In the process, we use `Selenium` [28] to automatically scrape all the search results and save the unique URL of each article. Next, `Selenium` loads the URLs and opens the `html` source code of each news story. We then rely on `BeautifulSoup` [29] to parse and clean the content of the news articles and save the resulting files in `.json` format. This process generated a collection of 158,514 articles.

B. Machine Learning classifier

Despite using Boolean filters in the EMIS query, the scraped articles include a considerable number of not relevant stories. To identify the specific news articles that are relevant to the study, we implement a ML classifier. Three human coders tagged as accept/reject a sample of 30,842 articles as initial training data. The accepted articles contain mentions of MCO-related incidents such as confrontations between criminals, drug seizures, or the capture of high-profile targets. The human coders marked as irrelevant articles not making direct reference to MCO-related incidents, containing editorial opinions about criminal violence, or cumulative security reports from authorities or journalists. The human coders also classified an additional random sample of 1,000 articles to assess their intercoder reliability, reaching an intercoder agreement of 90.4% and a Fleiss' Kappa of 0.704.

The initial training data contains an unbalanced acceptance rate of 23%. To balance this training data, we incorporate news articles from a manually collected data set [24] containing relevant articles for the same time period. This increased the sample to 60,837 articles with 61% of them categorized as accept. The next step consists in normalizing the text and eliminating diacritic characters, digits, punctuation marks, and stop words. We use the `SpaCy` [30] Spanish language lemmatizer to reduce words to their lemma, which facilitates standardizing the features dictionary on the reduced form of each word. Then, `TfidfVectorizer` from `sci-kit learn` [31] converts the raw data into a features matrix capped at 5,000 features. The pipeline shuffles and splits the training data into 5 folds, evaluates each model using k-fold cross validation, and assigns 10% of data for testing.

Fig. 2 reports the performance of different models based on F1 scores. We use a wide range of algorithms, from traditional approaches, ensemble methods, to deep learning. There are three Convolutional Neural Network (CNN) models reported from a random grid search, with a shared vocabulary size of 85,178 and an embedding dimension of 50. We include transformer models using the `simple transformers` library [32] and the `hugging face` transformers library [33]. The Extreme Gradient Boosting (XGB) model reports the lowest F1 average of 0.902. ALBERT is run using the `albert-base-v1` model and averages 0.914 F1 across

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

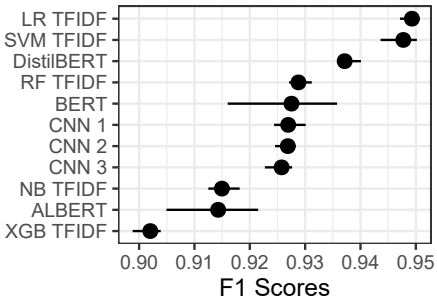


Fig. 2: Machine learning model performance.

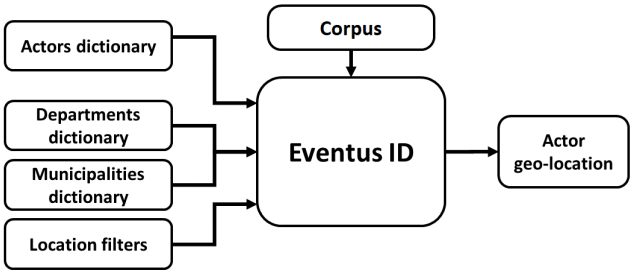


Fig. 3: Eventus ID system

folds. The Multinomial Naive Bayes (NB) model has an average F1 score of .915 across folds. CNN 3 uses 128 filters and a kernel size of 7, it averages 0.926 F1. CNN 2 has 64 filters and a kernel of size 7, and has an average of 0.9268 F1. CNN 1 produces marginally better results averaging 0.9269 with 128 filters and a kernel of size 3. BERT uses the `bert-base-multilingual-uncased` model with 5-fold cross validation averages an F1 of 0.928. The Random Forest Classifier (RF) generates an average F1 of 0.928. The best performing transformer is the DistilBERT using `distilbert-base-multilingual-cased`, it has an average of 0.937 F1 across folds. The Support Vector Machine (SVM) model averages an F1 of 0.947. The Logistic Regression (LR) model is the best performing model with an average F1 of 0.949 and it is the model used in the implementation stage.

The 158,514 collection of news articles use the same normalization and preprocessing pipeline as the training data. The LR model is then applied to this universe of articles, resulting in 43,681 stories on MCO related events. The human coders briefly reviewed a small sample of the classified articles and confirmed adequate performance.

C. Event coding

To generate the geo-referenced data on the territorial presence of Criminal Organizations in Mexico, we used Eventus ID [25], a supervised NLP application for event extraction from text written in Spanish. Eventus ID comes from a family of sparse parsing coders that started with Tabari [34], one of the first programs used for analyzing conflict data that ignited a vast research agenda [2], [7], [35]–[39]. This stream of coders eventually evolved into Petrarch [40], a new generation of event coding based on universal dependencies.

In general, event data is defined as a discrete description of someone doing something to someone else in a given time and place based on explicit information mentioned in the text. More technically, an event comprises five generic elements: the `source` is the actor conducting an action, the `action` being undertaken, the `target` of such action, a certain `date`, and a specific `location`. In this particular application, we customized Eventus ID's configuration to only code the actor, the location, and the date. The codification task focuses on geo-locating the presence of specific criminal groups in a given location without considering other behavioral characteristics contained in the text. In this way, coding the geographic location of actors is a simpler task than coding the full event.

To geo-reference the presence of Criminal Organizations, Eventus ID relies on three inputs: the corpus, the dictionary of actors, and the dictionaries of locations. Fig. 3 presents the intuition behind Eventus ID [25]. The algorithm reads each paragraph of the corpus. Eventus ID uses the entities listed in the dictionary of actors as search criteria and looks for the actors as mentioned in the corpus. Once the program identifies an actor, it uses the dictionaries of locations (at the state and municipal level) to look for a matching toponyms (names of places) indicating the location of occurrence. Finally, to minimize geographic ambiguity and identification of false locations, Eventus ID uses the filters dictionary.

For this application, the 2010-2018 part of the corpus comes from the collection of news reports discussed in the ML section. The earlier part of the corpus from 2000-2010 comes from Osorio [24]. The integrated corpus contains narratives of criminal activity in Mexico between 2000 and 2018. After cleaning and reformatting the text, the coding task processed a corpus of 230 MB of text.

To detect MCOs, Eventus ID relies on an actors dictionary containing a list of 5,797 entities including names of main criminal organizations, smaller criminal groups, cartel leaders, and key members. After developing an initial list of actors known to the researchers, we used Named Entity Recognition (NER) [41] to assist the expansion of the actors dictionary. The dictionary of actors contains information related to 10 main MCOs as reported in Table I. In addition, the dictionaries include generic mentions of unspecified criminal groups and specific names of minor criminal organizations as separate categories. In total, the dictionary allows classifying more than 200 criminal cells that can be clustered into the main MCOs mentioned above or analyzed in a disaggregated manner.

TABLE I: Main Mexican Criminal Organizations

Criminal Organizations	
Sinaloa Cartel	Michoacan Family
Juarez Cartel	Los Zetas
Gulf Cartel	Beltran-Leyva Cartel
Tijuana Cartel	La Barbie Cartel
New Generation Jalisco Cartel	Huachicoleros

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Finally, to detect MCO's locations, Eventus ID uses dictionaries of states and municipalities. In contrast to probabilistic methods for geo-referencing event data [17], [18], [42], Eventus ID relies on exact matching between the location dictionaries and the corpus content. These dictionaries include the names of all states and municipalities in the country, as well as redundancies such as abbreviations and different ways of spelling location names. After identifying an actor in the corpus, Eventus ID uses these dictionaries of locations to look for a matching toponym in the text indicating the territorial presence of the identified MCO.

After detecting the possible location of a criminal group, Eventus ID uses the filters dictionary to verify that such place is indeed a physical location and not a false geographic reference. The filters dictionary contains about 42,000 entries that prevent geographic ambiguity and reduce the risk of false positives. As an example to illustrate the function of this feature, the locations filter includes the entry "Juarez Cartel," a prominent MCO named after City of Juarez in the state of Chihuahua, where it has its headquarters. Including this entry in the filters dictionary helps to avoid erroneously coding the word "Juarez" as a geographic location when the algorithm matches the entity "Juarez Cartel" in the text.

Finally, the event coding output goes through a process of data cleaning, validation, and de-duplication to generate a database of the territorial presence of criminal organizations at the daily-municipality level. The final database adopts a conservative approach keeping only those records in which there is an explicit mention of a specific criminal organizations and it is possible to geo-referene the data at the daily municipal level. This implies discarding a considerable amount of information that lacks sufficient precision to geo-reference the violent presence of MCOs at such high degree of granularity.

To illustrate the functionality of the Eventus ID algorithm described above, consider the following example. Let the dictionary of actors contain the following names of criminal groups and their corresponding codes `actors={Los Zetas [607], Tijuana Cartel [604]}`. Consider also dictionaries of locations such that `state = {Baja California [2]}` and `municipality = {Mexicali [2002]}`. Finally, consider a filter of locations such that `filter = {Tijuana Cartel}`. Using these parameters, Eventus ID codes the following sentence as described below:

English: `Doc_20071003_XX_01 | Hit-men of Los Zetas had an armed confrontation with members of the Tijuana Cartel in Mexicali, Baja California.`

Spanish: `Doc_20071003_XX_01 | Sicarios de Los Zetas se enfrentaron con miembros del Cartel de Tijuana en Mexicali, Baja California.`

Output:
```
10/03/2007 607 2002 3
10/03/2007 604 2002 3
```

Eventus ID extracts the date from the file name as `10/03/2007`; it also identifies the Zetas and Tijuana Cartel as the relevant actors and assigns their codes `607, 604`; it also identifies their location in Mexicali, Baja Californaia and assigns their corresponding codes `2002 3`. Finally, the location filter will correctly identify that the Tijuana Cartel name refers to the criminal organization and not the municipality of Tijauna, which is also in the state of Baja California.

Implementing the full NLP process using Eventus ID generated more than 163,000 records of geo-referenced MCO activity at the daily-municipal level between 2000 and 2018. Finally, after identifying the geographic location of each actors, we assign the latitude and longitude geographic coordinates of each data point to facilitate the spatial visualization in a web interactive map.

D. Temporal Trends

To identify behavioral patterns, the study relies on different data visualization techniques including time series analysis and Geographic Information Systems (GIS). Fig. 4 presents the nationally-aggregated trends of the most prominent MCOs between 2000-2018. This graph presents a measure of intensity counting the total number of daily detections of each MCO in a municipality-year. The graph includes the Juarez Cartel, the Beltran-Leyva group, the Sinaloa Cartel, the Tijuana Cartel, the Gulf Cartel, the Barbie group, Los Zetas, the New Generation Jalisco Cartel (CJNG), the Michoacan Family, and oil thieves known as "Huachicoleros". As Fig. 4 shows, MCOs conducted two waves of dramatic expansion during the period of observation. The first wave took place between 2007-2011 and the second weave started in 2013 and runs up to 2018.

In addition to these aggregate trends, Fig. 5 presents the temporal trends of the four most aggressive organizations that largely contributed to the escalation of criminal violence in Mexico. Panel (a) in Fig. 5 shows the dramatic intensification of Los Zetas in two waves. Los Zetas is a highly sophisticated criminal group formed by elite troops from the Mexican Army that defected the government ranks and joined the Gulf Cartel as its enforcement branch [43] [44]. After the extradition of the leader of the Gulf Cartel, Los Zetas broke off and created their own independent organization. Los Zetas aggressively expanded their territory using a "franchise system" in which they coerced a multitude of small local criminal organizations

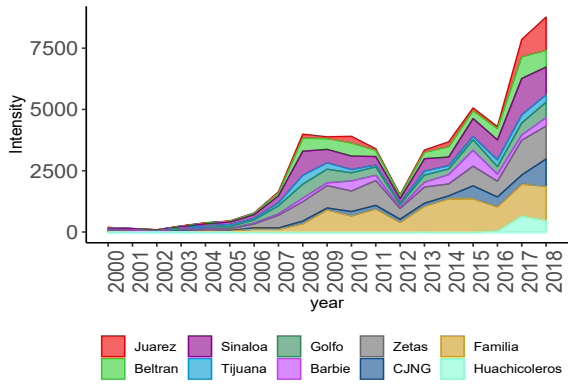
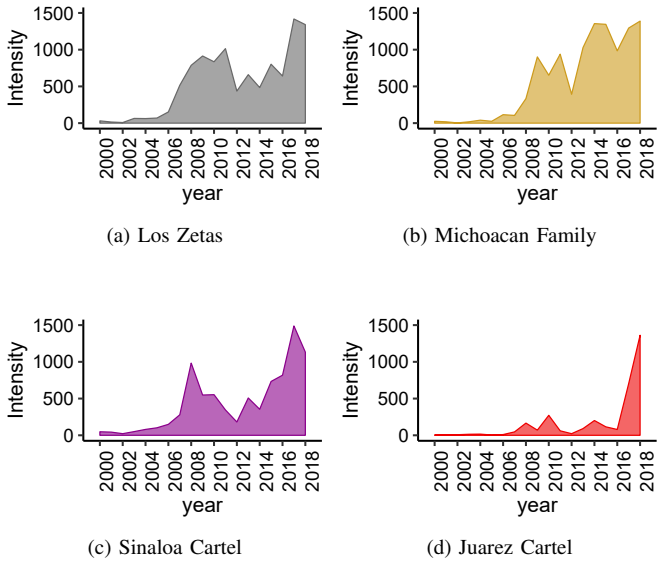


Fig. 4: Temporal trends of Mexican criminal organizations

978-1-7281-6026-2/20/$31.00 ©2020 IEEE



(a) Los Zetas (b) Michoacan Family

(c) Sinaloa Cartel (d) Juarez Cartel

Fig. 5: Main criminal organizations in Mexico

and allowed them to carry their name [45]. This rapid escalation explains the first wave of territorial expansion between 2007 and 2011 in Panel (a). After government officials arrested or killed several Zeta leaders, the organization declined and fragmented. However, starting in 2015 a new wave of smaller groups formerly associated with Los Zetas[1] carried out a second expansive wave.

Another very aggressive MCO is the Michoacan Family. As Panel (b) in Fig. 5 shows, the Michoacan Family rapidly expanded its territory after the Mexican government inaugurated the war on drugs in December 2006 by cracking down on this MCO. Government authorities concentrated their attention in neutralizing the leadership of this group and managed to undermine it in 2012. However, the main Michoacan Family organization fragmented into a variety of smaller groups[2] that engaged in more intense levels of violence starting in 2013.

Panel (c) in Fig. 5 shows the time series of intensity of the Sinaloa Cartel. This criminal organization has its stronghold in the Western highlands of Mexico in areas long associated with poppy and marijuana cultivation, and along traditional trade routes along the Pacific coast leading up to the Mexico-U.S. border [46] [47]. Under the leadership of "El Chapo" Guzman, the Sinaloa Cartel carried out an aggressive military campaign to secure entry points to the U.S. along the border. Starting in 2007, the Sinaloa Cartel led incursions into the Tijuana Cartel in the Northern state of Baja Californa and against the Juarez Cartel in the state of Chihuahua. The first wave of intensification in Panel (c) between 2007 and 2010 corresponds to this expansionist effort. During this period, the Sinaloa Cartel suffered two major internal fractures that led to the creation of the Belran-Leyva organization and the criminal group of La Barbie (see the green and pink trends in Fig. 4, respectively). The Sinaloa Cartel carried out a second expansion wave in 2014, when government authorities arrested "El Chapo." The arrest weakened the position of the Sinaloa Cartel and softened its grip on the territories of the Tijuana and Juarez cartels that invaded a few years earlier. This weakening, led to increasing contestation from rival cartels and eroded discipline within the organization, thus opening the door to internal disputes [48]. "El Chapo" managed to escape from prison, but was re-apprehended in 2016 and quickly extradited to the United States. The effective removal of the Sinaloa Cartel leadership favored increasing contestation from competing cartels and internal factions.

Finally, Panel (d) in Fig. 5 presents the temporal trends of the Juarez Cartel. This is an old criminal organization operating in the border town of Juarez, Chihuahua since the 1970s [49], which is a strategic entry point to the U.S. drug market and has long been a valuable territory. The Juarez Cartel intensified its violent activities in 2008 and reached a first peak in 2010. This escalation of violence was a both a response against law enforcement operations and against the incursion of the rival Sinaloa Cartel that was trying to expand into its territory. During this period, the Juarez Cartel boosted its military capabilities by creating "La Línea", a group of specialized hit-men in charge of defending its territory, and later on "Los Aztecas". The Juarez Cartel launched a second aggressive escalation in 2016, reaching an unprecedented peak of intensity in 2018. This second wave of activity corresponds to the period of extradition to the U.S. of "El Chapo" Guzman, the leader of the rival Sinaloa Cartel that tried to seize control over Juarez for several years. It seems that the fall of "El Chapo" opened the opportunity for the Juarez Cartel to fight back in an effort to expel the intruders.

E. Spatial Trends

In addition to disaggregating the temporal trends, the fine-grained resolution of the data allows to identify the spatial dynamics of cartel territories. Based on geo-referenced data at the municipality level, the data visualization relies on Geographic Information Systems (GIS) tools to analyze the spatial trends. Using ArcGIS Pro and ArcGIS Online, we present an interactive web map that renders heat-maps of main areas of MCO concentration. The application is available at www.ocved.mx. This interface allows users to visualize the territorial presence of criminal organizations by aggregating all MCOs or to track the areas of operation of specific criminal groups. The timer also allows to display the territorial dynamics of concentration and expansion of cartel presence over time by aggregating data on a yearly basis. In addition, the dynamic rendering of the application recalculates the heat-maps as the user zooms in or out the scale of the map.

[1]These include: Los Zetas, Los Hijos del Diablo, Los Enrique, Los Guerreros, Los Numeros, Los Enfermeros, Cartel del Noreste, Los Cotorros, Los Lancheros, Los Broncos, and Los Legendarios.

[2]These include: La Familia Michoacana, Caballeros Templarios, Cartel de Tlahuac, La Empresa, Los Troyanos , La Nueva Empresa, Champis, Brown Side Family, Los Gordos, Los Perez, Los Pumas, El Charro, Guardia Morolense, Los Viagras, Los Jaguares, La Nueva Familia Michoacana, Cartel de Tepalcatepec, Cartel de Zicurian, and Los Tequileros.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

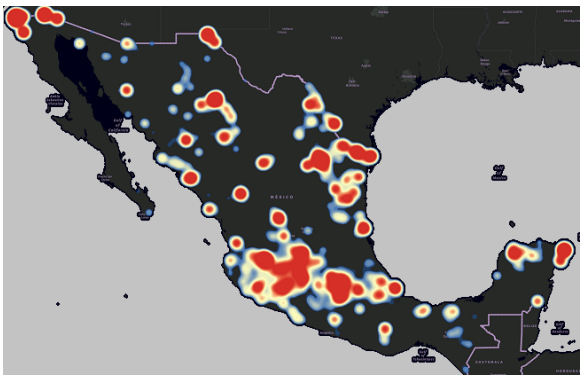
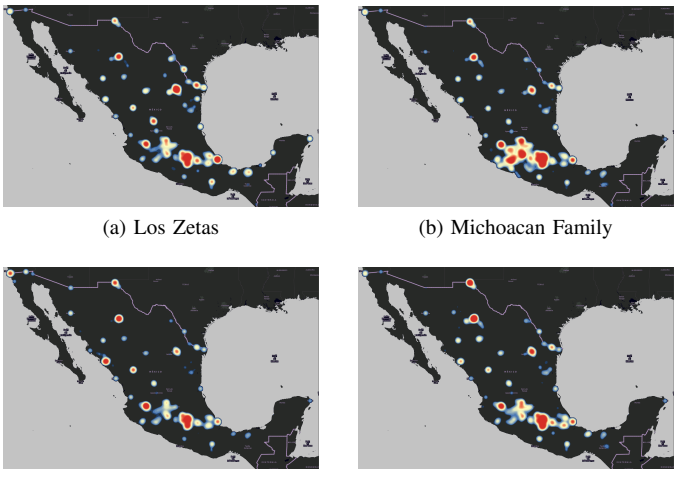


Fig. 6: Spatial distribution of criminal organizations

To illustrate these spatial trends, Fig. 6 presents the overall territorial concentration of Criminal Organizations in Mexico between 2000 and 2018. This aggregate level data shows that a vast part of the country suffers the presence of criminal organizations. The data also shows that such presence is not evenly distributed as there are several areas of intense criminal activity along the U.S.-Mexico border, as well as along the Pacific Coast and the Gulf of Mexico, and in the highlands that are favorable for illicit crop cultivation.

The granularity of the geo-referenced data used in this application, allows users of the interactive web interface to disaggregate the data in order to identify the territorial dynamics of specific criminal groups. Fig. 7 presents a series of maps displaying the territories of some MCOs showing distinct areas of concentration. For example, Panel (a) in Fig. 7 shows the areas of operation of Los Zetas, Panel (b) presents the Michoacan Family territory, Panel (c) indicates the geographic concentration of the Sinaloa Cartel, and Panel (d) shows the territorial reach of the Juarez Cartel.

In general, the granularity of the data allows identifying the temporal and spatial trends of different MCOs. The distinct trajectories of each group challenge the notion of criminal organizations behave in a homogeneous manner and reveals their dynamism and heterogeneity. Such information provides a solid empirical foundation for advancing substantive research on the dynamics of drug-related violence, territorial competition of criminal organizations, and the emergence of criminal governance, as well as their consequences on the political and economic realms [24] [50] [51] [52] [53] [54] [55].



(a) Los Zetas (b) Michoacan Family

(c) Sinaloa Cartel (d) Juarez Cartel

Fig. 7: Specific cartel territories

IV. Conclusion

In contrast to existing approaches for detecting criminal organizations in Mexico, this article relies on a larger universe of articles for a broader time period, and implements ML and NLP tools to generate accurate and disaggregated data by day and location. The tools developed for scraping and classifying news articles may serve in future applications that continuously update the presence of MCOs using text-as-data. Our trained model outperforms existing implementations of ML in classifying MCO related news, and the use of an LR model allows for less computationally expensive applications. The performance can be attributed to our use of multiple human coders and the sheer volume of our training data.

Using NLP tools such as Eventus ID enables extracting geo-coded information of criminal groups at the daily-level in a systematic and reliable manner. The dictionary of actors used in this application is robust and includes an encompassing list of the most relevant criminals and their organizations through the end of 2018. Location filters incorporated into Eventus ID also offer an advantage of geographic disambiguation as compared to existing data sets in that it discriminates false positives based on a dictionary of contextualized information.

The visual representation helps grasp the significance of our contributions to analyzing organized crime. The temporal analysis reveals the heterogeneous fluctuations of violent activity of different criminal groups. In addition, the interactive map helps to identify dynamics of territorial expansion and contraction of cartel presence in Mexico. Overall, the NL and NLP applications presented in this study provide solid empirical foundations to advance substantive academic and policy analysis to better understand and control organized criminal activity in Mexico and in other latitudes.

References

[1] P. A. Schrodt and D. Van Brackel, "Automated coding of political event data," in *Handbook of Computational Approaches to Counterterrorism*, D. Subramanian, Ed. New York: Springer, 2013, pp. 23–50.

[2] S. Chojnacki, C. Ickler, M. Spies, and J. Wiesel, "Event data on armed conflict and security: New perspectives, old challenges, and some solutions," *International Interactions*, vol. 38, no. 4, pp. 382–401, 2012.

[3] P. A. Schrodt, B. Stewart, J. Lautenschlager, A. Shilliday, D. V. Brackel, and W. Lowe, "Automated production of high-volume, near-real-time political event data," 2010, unpublished.

[4] T. B. Seybolt, J. D. Aronson, and B. Fischhoff, Eds., *Counting Civilian Casualties: An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict*. Oxford: Oxford University Press, 2013.

[5] C. Davenport and P. Ball, "Views to a kill: exploring the implications of source selection in the case of Guatemalan State Terror, 1977-1995." *Journal of Conflict Resolution*, vol. 46, no. 3, pp. 427–450, 2002.

[6] Y. M. Zhukov, C. Davenport, and N. Kostyuk, "Introducing xSub: A new portal for cross-national data on subnational violence," *Journal of Peace Research*, vol. 56, no. 4, pp. 604–614, 2019.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

[7] J. Hammond and N. B. Weidmann, "Using machine-coded event data for the micro-level study of political violence," *Research and Politics*, vol. 1, no. 2, pp. 1–8, 2014. [Online]. Available: http://rap.sagepub.com/content/1/2/2053168014539924.full#ref-11

[8] A. Hanna, "Developing a system for the automated coding of protest event data," 2014, unpublished. [Online]. Available: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2425232

[9] M. Ward, A. Beger, J. Cutler, M. Dickenson, C. Dorff, and B. Radford, "Comparing GDELT and ICEWS event data," 2013. [Online]. Available: https://benradford.github.io/images/publications/GDELTICEWS.pdf

[10] B. E. Bagozzi, P. T. Brandt, J. R. Freeman, J. S. Holmes, A. Kim, A. Palao Mendizabal, and C. Potz-Nielsen, "The prevalence and severity of underreporting bias in machine and human-coded data," *Political Science Research and Methods*, pp. 1–9, mar 2018. [Online]. Available: https://www.cambridge.org/core/product/identifier/S2049847018000110/type/journal_article

[11] N. B. Weidmann, "On the accuracy of media-based conflict event data," *Journal of Conflict Resolution*, vol. 59, no. 6, pp. 1129–1149, 2015. [Online]. Available: https://doi.org/10.1177/0022002714530431

[12] K. Donnay, E. T. Dunford, E. C. McGrath, D. Backer, and D. E. Cunningham, "Integrating conflict event data," *Journal of Conflict Resolution*, vol. 63, no. 5, pp. 1337–1364, 2019.

[13] K. Eck, "In data we trust? A comparison of UCDP GED and ACLED conflict events datasets," *Cooperation and Conflict*, vol. 47, no. 1, pp. 124–141, 2012.

[14] J. Grimmer and B. M. Steward, "Text as data: The promise and pitfalls of automatic content analysis methods for political texts," *Political Analysis*, vol. 21, no. 3, pp. 267–297, 2013. [Online]. Available: http://www.jstor.org/stable/24572662

[15] J. Wilkerson and A. Casas, "Large-scale computerized text analysis in political science: Opportunities and challenges," *Annual Review of Political Science*, vol. 20, no. 1, pp. 529–544, 2017. [Online]. Available: https://doi.org/10.1146/annurev-polisci-052615-025542

[16] G. E. Hollibaugh, "The use of text as data methods in public administration: A review and an application to agency priorities," *Journal of Public Administration Research and Theory*, vol. 29, no. 3, pp. 474–490, 2018.

[17] S. Lee, H. Liu, and M. Ward, "Lost in space: Geolocation in event data," *Political Science Research and Methods*, vol. 7, no. 4, pp. 871–888, 2019. [Online]. Available: https://arxiv.org/abs/1611.04837

[18] A. Halterman, "Geolocating political events in text," 2019. [Online]. Available: https://arxiv.org/pdf/1905.12713

[19] J. Osorio, M. Mohamed, V. Pavon, and S. Brewer-Osorio, "Mapping violent presence of armed actors in Colombia," *Advances of Cartography and GIScience of the International Cartographic Association*, 2019.

[20] C. Raleigh, A. Linke, H. Hegre, and J. Karlsen, "Introducing ACLED: an armed conflict location and event dataset: special data feature," *Journal of peace research*, vol. 47, no. 5, pp. 651–660, 2010.

[21] R. Sundberg and E. Melander, "Introducing the UCDP georeferenced event dataset," *Journal of Peace Research*, vol. 50, no. 4, pp. 523–532, 2013.

[22] E. Boschee, J. Lautenschlager, S. O'Brien, S. Shellman, J. Starz, and M. Ward, "ICEWS coded event data," 2016. [Online]. Available: https://dx.doi.org/10.7910/DVN/28075

[23] S. Althaus, J. Bajjalieh, J. F. Carter, B. Peyton, and D. A. Shalmon, "Cline Center Historical Phoenix Event Data," 2019. [Online]. Available: https://doi.org/10.13012/B2IDB-0647142{_}V2

[24] J. Osorio, "The contagion of drug violence: spatiotemporal dynamics of the Mexican war on drugs," *Journal of Conflict Resolution*, vol. 59, no. 8, pp. 1403–1432, 2015. [Online]. Available: https://doi.org/10.1177/0022002715587048

[25] J. Osorio and A. Reyes, "Supervised event coding from text written in Spanish: Introducing Eventus ID," *Social Science Computer Review*, vol. 35, no. 3, pp. 406–416, 2017. [Online]. Available: https://doi.org/10.1177/0894439315625475

[26] M. Coscia and V. Rios, "Knowing where and how criminal organizations operate using web content," in *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2012, pp. 1412–1421. [Online]. Available: https://dl.acm.org/doi/10.1145/2396761.2398446

[27] F. Sobrino, "Mexican cartel wars: Fighting for the US opioid market," unpublished. [Online]. Available: https://www.fersobrino.com/files/DraftPaper.pdf

[28] Selenium, "Selenium webdriver," *Selenium HQ*, 2013. [Online]. Available: https://selenium.dev/documentation/en/webdriver/

[29] L. Richardson, "Beautiful soup documentation," *April*, 2007. [Online]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[30] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017, to appear.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in Python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[32] T. Rajapakse, "Simple transformers," May 2020. [Online]. Available: https://simpletransformers.ai/

[33] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Huggingface's transformers: State-of-the-art natural language processing," 2019.

[34] P. A. Schrodt, "TABARI. textual analysis by augmented replacement instructions," Lawrence, Kansas, 2009. [Online]. Available: http://eventdata.parusanalytics.com/software.dir/tabari.html

[35] R. H. Best, C. Carpino, and M. J. C. Crescenzi, "An analysis of the TABARI coding system," *Conflict Management and Peace Science*, vol. 30, no. 4, pp. 335–348, jul 2013.

[36] P. A. Schrodt, "Automated coding of international event data using sparse parsing techniques." Chicago: Paper presented at the annual meeting of the International Studies Association, 2001. [Online]. Available: http://polmeth.wustl.edu/media/Paper/schro01b.pdf

[37] P. A. Schrodt, D. Gerner, and O. Yilmaz, "Using event data to monitor contemporary conflict in the Israeli-Palestine dyad," in *Annual Meeting of the International Studies Association*, 2004.

[38] P. A. Schrodt, "Twenty years of the Kansas event data system project," *The Political Methodologist*, vol. 14, no. 1, pp. 2–6, 2006.

[39] P. A. Schrodt and D. Gerner, "Fundamentals of machine coding," in *Analyzing International Event Data: A Handbook of Computer-Based Techniques*, 2012. [Online]. Available: http://parusanalytics.com/eventdata/papers.dir/AIED.Preface.pdf

[40] P. A. Schrodt, J. Beieler, and M. Idris, "Three's a charm?: Open event data coding with EL:DIABLO, PETRARCH, and the Open Event Data Alliance," in *International Studies Association*, Toronto, 2014. [Online]. Available: http://parusanalytics.com/eventdata/papers.dir/Schrodt-Beieler-Idris-ISA14.pdf

[41] The Stanford Natural Language Processing Group, "Stanford named entity recognizer," Palo Alto, CA, 2014. [Online]. Available: http://nlp.stanford.edu/software/CRF-NER.shtml

[42] M. B. Imani, S. Chandra, S. Ma, L. Khan, and B. Thuraisingham, "Focus location extraction from political news reports with bias correction," in *2017 IEEE International Conference on Big Data (Big Data)*. Boston, MA: Institute of Electiracal and Electronics Engineers, 2017. [Online]. Available: 10.1109/BigData.2017.8258141

[43] R. Ravelo, *Crónicas de Sangre. Cinco Historias de los Zetas*. Mexico City: Debolsillo, 2007.

[44] D. E. Osorno, *La Guerra de los Zetas*. Mexico: Grijalbo, 2012.

[45] J. Bailey, ""Los Zetas" y McDonald's," nov 2011. [Online]. Available: http://www.eluniversal.com.mx/editoriales/55520.html

[46] L. Astorga, *Drogas Sin Frontera*. Mexico City: Grijalbo, 2003.

[47] ——, *El siglo de las drogas: el narcotráfico, del porfiriato al nuevo milenio*. Mexico City: Plaza y Janés, 2005.

[48] P. H. Reuter, "Systemic violence in drug markets," *Crime, Law and Social Change*, vol. 52, no. 3, pp. 275–284, 2009.

[49] F. Cruz, *El Cártel de Juárez*. Mexico City: Planeta, 2008.

[50] R. Snyder and A. Durán-Martínez, "Drugs, violence and state-sponsored protection rackets in Mexico and Colombia," *Colombia Internacional (on-line)*, vol. 70, no. July/December, pp. 61–91, 2009.

[51] A. Duran-Martinez, "Criminals, cops, and politicians: Dynamics of drug violence in Colombia and Mexico," Providence, 2013.

[52] M. Dell, "Trafficking networks and the Mexican drug war," *American Economic Review*, vol. 105, no. 6, pp. 1738–1779, 2015.

[53] N. Barnes, "Criminal politics: An integrated approach to the study of organized crime, politics, and violence," *Perspectives on Politics*, vol. 15, no. 4, pp. 967–987, 2017.

[54] D. E. Arias, *Criminal Enterprises and Governance in Latin America and the Caribbean*. New York: Cambridge University Press, 2017.

[55] B. Lessing and G. D. Wills, "Legitimacy in criminal governance: Managing a drug empire from behind bars," *American Political Science Review*, vol. 113, no. 2, pp. 584–606, 2019.