Classifying Organized Criminal Violence in Mexico using ML and LLMs

Javier Osorio
School of Government and Public Policy
University of Arizona
josorio1@arizona.edu

Juan Vásquez
Department of Computer Science
University of Colorado Boulder
juan.vasquez-1@colorado.edu

Abstract

Natural Language Processing (NLP) tools have been rapidly adopted in political science for the study of conflict and violence. In this paper, we present an application to analyze various lethal and non-lethal events conducted by organized criminal groups and state forces in Mexico. Based on a large corpus of news articles in Spanish and a set of high-quality annotations, the application evaluates different Machine Learning (ML) algorithms and Large Language Models (LLMs) to classify documents and individual sentences, and to identify specific behaviors related to organized criminal violence and law enforcement efforts. Our experiments support the growing evidence that BERT-like models achieve outstanding classification performance for the study of organized crime. This application amplifies the capacity of conflict scholars to provide valuable information related to important security challenges in the developing world.

1 Introduction

Recent advancements in Natural Language Processing (NLP) have revolutionized political science analyses by enabling efficient and accurate analysis of large volumes of text. These tools have demonstrated impressive capabilities in tackling complex text analysis tasks, leading to their increasing adoption by political scientists including conflict scholars specialized in the study of violence and crime (Hu et al., 2022; Halterman et al., 2023b,a; Hürriyetoğlu et al., 2022; Motlicek, 2023). In this paper, we present an application of various Machine Learning (ML) algorithms and Large Language Models (LLMs) to analyze a variety of behaviors by organized criminal groups in Mexico by processing text written in Spanish.

By leveraging these state-of-the-art NLP techniques, we aim to make significant contributions to the study of organized criminal violence and law enforcement efforts in developing countries. Based on a set of high-quality annotations, we train and evaluate different ML algorithms and LLMs to determine their effectiveness in detecting and categorizing organized crime violence and law enforcement. Furthermore, we extend our analysis beyond simple document-level classification by analyzing the relevance of specific sentences within documents and then analyzing the specific types of events described in the narratives. In this way, we move beyond the identification of organized criminal groups as named entities (Osorio and Beltrán, 2020; Coscia and Rios, 2012; Signoret et al., 2021), and focus on analyzing criminal behaviors.

Results show the high levels of performance of BERT-like models to effectively classify relevant news articles, as well as relevant sentences within them. The models also have remarkable results for classifying a variety of violent and non-violent actions perpetrated by criminal groups or conducted by law enforcement forces.

Overall, our research emphasizes the advantages of leveraging NLP tools in political science research, particularly in the domain of political violence and organized crime analysis. By exploiting their remarkable capabilities for document, sentence, and class classification, researchers can extract valuable insights from vast corpora in local languages, thus enabling a more comprehensive understanding of complex social behaviors. The elements advanced in this research can pave the way toward the development of a fully integrated ML crime analysis system in Spanish.

2 Recent Developments

NLP researchers have advanced various supervised learning and deep-learning architectures to address a variety of text analysis challenges (Thangaraj and Sivakami, 2018; Minaee et al., 2021). Due to the complexities of analyzing unstructured text, rule-based developments showed limited performance when tackling complex NLP tasks until the emer-

Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, pages 1–10, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-089-2_001

gence of pre-trained language models. In particular, Google's Bidirectional Encoder Representations from Transformers (BERT) language model (Devlin et al., 2019) became a game-changer in a variety of NLP tasks. After BERT's initial development in English, Google released multilingual BERT (mBERT). Political scientists quickly noted these NLP tools and applied them to a variety of tasks relevant to political analysis (Kowsari et al., 2019; Rodriguez and Spirling, 2022; Terechshenko et al., 2020; Lowe and Benoit, 2013; Rudkowsky et al., 2018; Häffner et al., 2023).

Computerized text analysis has a long trajectory in the study of international conflict, but recent ML developments are just gaining traction among conflict scholars. Early efforts to identify incidents of political conflict or cooperation using text analysis relied on complex systems of rules (Schrodt et al., 2010; Schrodt and Van Brackel, 2013; Boschee et al., 2016; Ward et al., 2013; Osorio and Reyes, 2017; Osorio et al., 2019). Unfortunately, these rule-based systems are too rigid and expensive to update, and the algorithms showed limited performance when tackling even basic NLP tasks.

Due to the limitations of rule-based approaches, recent NLP developments such as ConfliBERT (Hu et al., 2022) and POLECAT (Halterman et al., 2023b,a) bring more flexibility and effectiveness in analyzing political violence. Unfortunately, those tools are focused exclusively on the English language. To address this challenge, scholars such as Hürriyetoğlu et al. (2022), Caselli et al. (2021), and Yang et al. (2023) have been advancing multilingual ML tools and LLM to study conflict.

Within this constellation of research, scholars have been using NLP tools to study organized crime in Mexico by processing text written in Spanish. Early efforts relied on rule-based approaches to track the territorial presence of organized criminal groups (Osorio and Reyes, 2017; Osorio, 2015; Coscia and Rios, 2012; Signoret et al., 2021). A common limitation of these studies is their exclusive focus on tracking the location of criminal groups. Unfortunately, this only provides information about "who" is present but does not say much about their behavior. Although (Osorio and Beltrán, 2020) and (Parolin et al., 2021) have been incorporating ML approaches to study organized crime, these ML applications have only focused on a narrow set of behaviors. To address these limitations, this study provides a fully integrated ML application to identify a broad range of behavioral trends of criminal groups and state authorities from news stories written in Spanish.

3 Training Data

Computational social scientists have paid increasing attention to the quality of training data annotations (Grimmer and Stewart, 2013; Hsueh et al., 2009; Erlich et al., 2022; Krommyda et al., 2021). Due to the need for high-quality annotations to maximize ML performance, this study implements a rigorous annotation protocol. To generate the training data, the study relied on a group of three human annotators supervised by the Principal Investigator (PI). The meticulous training, supervision, and validation protocols implemented in this project allowed generating high-quality annotations. The protocol consisted of human annotators classifying information from high a level of aggregation to progressively fine-grained annotations in three stages: document classification (task 1), sentence relevance (task 2), and event type (task 3).

Task 1: Document relevance. To ensure the validity of the data at the highest level of aggregation, the first task consists of identifying news articles conveying information on organized criminal violence and law enforcement efforts against criminal groups. Failing to discriminate the domain-specificity of the documents increases the risk of including false positives which are likely to undermine the ML performance and the output validity. This study relies on the document classification originally conducted by Osorio and Beltrán (2020), who used a team of human annotators to classify news articles as "relevant" or "not relevant". The first step consisted of using a query to gather news articles from 110 national and local newspapers in Mexico. Then, annotators classified as "relevant" news reports that provide descriptions of factual incidents of criminal violence or law enforcement against criminal groups. These incidents include armed confrontations between criminals; armed clashes between criminals and government authorities; arrests of members of criminal groups; drug seizures; seizures of assets (e.g. vehicles, money, real state); seizures of weapons; or the capture of high-profile targets. The team of annotators classified as "not relevant" news stories that do not make direct reference to organized criminal violence events; editorial opinions about criminal violence; statements or claims from victims, civil

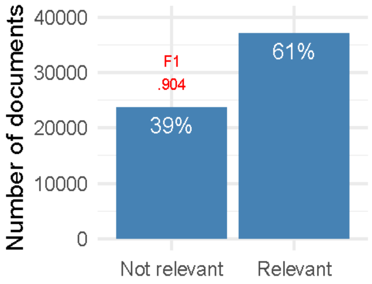


Figure 1: Annotations of relevant news articles (task 1)

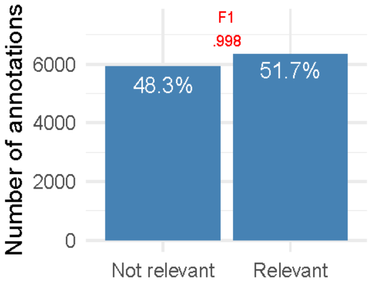


Figure 2: Annotations of relevant sentences (task 2)

society organizations, or government officials; or summaries from government authorities providing a cumulative report of law enforcement activities.

The training data at the document level consists of 60,837 news articles in Spanish, out of which 61% are "relevant" and the other 39% are "not relevant" (see Figure 1). This large training data was produced with high inter-annotator reliability (F1=0.904), reported by Osorio and Beltrán (2020).

Task 2: Sentence relevance. This study goes beyond the document-level classification initially implemented by Osorio and Beltrán (2020) and analyzes the relevance of specific sentences. To do so, we selected a random sample of relevant documents from task 1 and disaggregated them into individual sentences. A team of human annotators classified sentences as "relevant" or "not relevant" following the criteria proposed by Osorio and Beltrán (2020). An initial group of six annotators underwent a three-week training program to gain familiarity with the ontology. In this process, the annotators labeled the same corpus in several rounds. Then, the PI selected the three annotators with the highest inter-annotator reliability score (F1>0.8) to work on task 2.

We used `www.tagtog.com`, a web-based annotation platform, to annotate a collection of 12,252 sentences. Under the PI's supervision, the team of annotators independently classified each sentence and implemented a cross-validation process consisting of several rounds of revision to ensure the consistency of their labeling decisions. After each sentence received three validated classifications, the team generated the gold standard record (GSR). To do so, the PI randomly assigned a set of sentences to each annotator, who evaluated the set of anonymous annotations from the previous round and determined the most accurate one as the GSR. Figure 2 shows the binary annotation that produced a balanced collection of 51.7% of the sentences as "relevant" and the other 48.3% as "not relevant", with a high inter-annotator agreement of F1=0.9982.

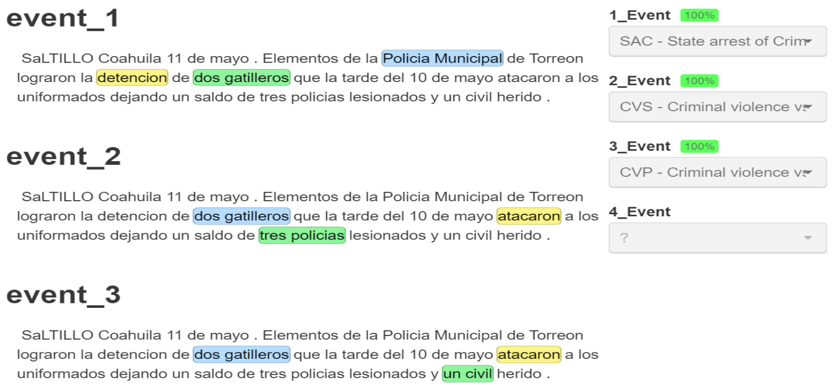
Task 3: Event type. The next step consists of annotating the type of event in the relevant paragraphs derived from the previous step. To do so, the annotators relied on a detailed codebook to classify 11 different types of events: (i) Criminal violence vs. criminals, (ii) Criminal violence vs. state, (iii) Criminal violence vs. civilians, (iv) Drug trafficking or production, (v) State violence vs. criminals, (vi) State arrest of criminals, (vii) State seizure of drugs, (viii) State seizure of guns, (ix) State seizure of assets, (x) State violence vs. civilians, and (xi) Civilian violence vs. criminals.

Classifying unstructured text using a large number of categories can be challenging, particularly when the narrative conveys information about multiple events. Identifying multiple actors conducting different actions in the same sentence can generate intractable annotation schemes. To address this challenge, the PI modified the annotation space to enable multiple-actor-action classification. Figure 3 shows the interface using the same sentence to classify three different actions: an arrest (event 1), an attack on the police (event 2), and violence against civilians (event 3). The interface allows coding up to four distinct types of actions from the same sentence.[1]

According to the annotation output in Figure 4, about 51.7% of the sentences contain a single event, 12.2% sentences have two events, 3.9% include three events, and 1.3% contain four events. The inter-annotator reliability assessment also indicates a high level of agreement between annotators with an F1=0.998 in event 1, F1=0.997 in event 2, and F1=1 in both events 3 and 4.

The team of annotators classified a total of 8,466

[1] Note that the interface in Figure 3 also allows annotating the span of text, a task that will be explored in future work for fine-grained text extraction and Named Entity Recognition.



event_1

SaLTILLO Coahuila 11 de mayo . Elementos de la Policia Municipal de Torreon lograron la detencion de dos gatilleros que la tarde del 10 de mayo atacaron a los uniformados dejando un saldo de tres policias lesionados y un civil herido .

event_2

SaLTILLO Coahuila 11 de mayo . Elementos de la Policia Municipal de Torreon lograron la detencion de dos gatilleros que la tarde del 10 de mayo atacaron a los uniformados dejando un saldo de tres policias lesionados y un civil herido .

event_3

SaLTILLO Coahuila 11 de mayo . Elementos de la Policia Municipal de Torreon lograron la detencion de dos gatilleros que la tarde del 10 de mayo atacaron a los uniformados dejando un saldo de tres policias lesionados y un civil herido .

1_Event 100%
SAC - State arrest of Crim

2_Event 100%
CVS - Criminal violence vs

3_Event 100%
CVP - Criminal violence vs

4_Event
?

Figure 3: Annotation interface

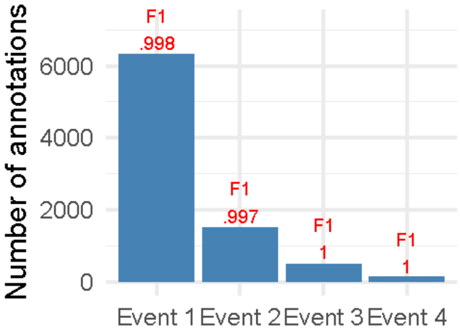


Figure 4: Event annotations

events. Figure 5 presents the distribution of annotations by different event categories and their perpetrators. Among the actions carried out by the state, the most frequent types of events are arrests (8.9%), followed by drug seizures (8.5%), gun seizures (7%), and seizures of assets (4%). Violence perpetrated by the state is rare. Annotators only identified state violence vs. criminals in 3.3% of the cases and violence vs. civilians in 0.2%.

Among the events initiated by criminal groups, criminal violence against civilians stands out as the most common category with 20.9% of the cases. In contrast, violence between criminal groups is less frequent (3.2%). In practice, it is difficult to distinguish between criminal violence against the population and against rival criminals using news articles. The reason is that news reports tend to provide a generic description of the incident without giving details about the victims. For example, an article may just indicate that "a group of hitmen conducted an attack and killed three men." In this case, the criminal character of the perpetrators is clear, but there is no information about the victims. The annotation protocol used in this research classifies this type of event as violence against the population.[2] Following the codebook, annotators classified events of violence between criminals when the news reports explicitly mention the criminal character of the perpetrator as well as the victim (e.g. name of the criminal group or the person's role such as hitman or lieutenant).

Criminal violence against the state constitutes the second most frequent annotation of criminal behavior (9.4%). In addition, annotators detected instances of criminal violence against the state in 3.3% of the sentences. Finally, the annotation output indicates that violence from the population against criminals is very rare, with only 0.2%.

Overall, as Figure 5 shows, the distribution of annotations in the training data is not balanced. There are some categories with a substantial number of annotations (particularly, criminal violence against civilians), while others do not have many annotations. This could affect the performance of ML algorithms and it is not plausible to expect good performance in categories with scarce annotations. Section 6.1 below discusses future research to address this challenge.

4 Experiment Setting

This study analyzes organized criminal violence in Mexico using a set of experiments to progressively process finer-grained information. The first stage focuses on classifying relevant documents. The second stage consists of classifying relevant sentences extracted from the relevant documents identified in the previous step. The final stage classifies the

[2] This coding decision rests on methodological and ethical grounds. Methodologically, annotators only classify information based on explicit evidence in the news report and make no assumptions about the victims. Ethically, the annotating procedure is based on the victim's presumption of innocence, which helps to reduce double victimization and stigmatization (Moon and Treviño-Rangel, 2020).

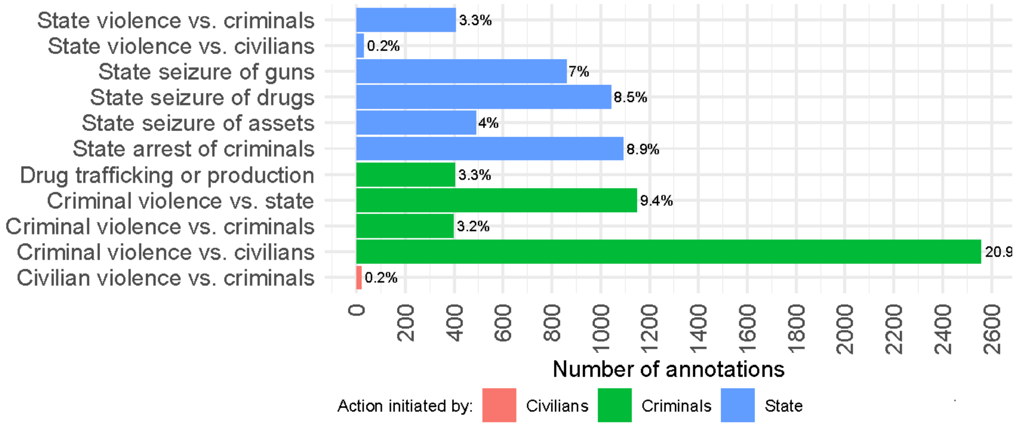


Figure 5: Annotations of event types (task 3)

specific type of organized criminal violence or law enforcement actions contained in the relevant sentences. This sequential approach helps to ensure the validity of the output data based on the concatenated focus on documents, sentences, and events. The large number and high quality of the annotations included in the training data provide a strong empirical foundation to assess the performance of the different ML algorithms and LLM considered.

Task 1: Document relevance. To address the challenge of determining which news articles are relevant to the topic of organized criminal violence, the study approaches this problem as a binary classification task at the document level. Based on the annotations provided in the training data, a positive outcome is operationalized as "relevant" and a negative outcome as "not relevant." In line with standards in computer science research, the experiment setup takes an agnostic approach and puts a variety of algorithms to compete in this binary classification task. This experiment considers a set of five traditional ML algorithms and three LLMs: Multinomial Naive Bayes (NB), Logistic Regression (LR), Random Forest Classifier (RF) (Breiman, 2001), Support Vector Machine (SVM) (Cortes and Vapnik, 1995), Extreme Gradient Boosting (XGB) (Chen and Guestrin, 2016), BETO (José et al., 2020), and Multilingual BERT (mBERT) (Devlin et al., 2018). The experimental setting uses this whole set of ML algorithms and LLM to assess the performance of tasks 2 and 3.

First, we pre-processed the data. We lemmatized the text in the corpus using the `es_core_news_sm` model from spaCy. Then, we removed Latin diacritics and stop-words. Next, we trained the traditional ML algorithms using their default settings in scikit-learn, and fine-tuned the LLMs using the Hugging Face library. We split the corpus into 90% for training and 10% for evaluation. To test the performance of our classifiers, we used the metrics implemented in scikit-learn.

Task 2: Sentence relevance. Based on the selection of relevant news stories derived from the previous stage, the experiment setup then focuses on classifying relevant sentences within the relevant documents. To do so, the study approaches this task as a binary classification at the sentence level. Based on the annotations, a positive outcome is operationalized as a "relevant" sentence, while a negative outcome indicates "not relevant" sentences. For the automatic classification, we follow the pipeline described for Task 1 and the experiments evaluate the full set of ML and LLM.

Task 3: Event type. The final set of experiments use a variety ML algorithms and LLM to classify different types of events at the sentence level. To do so, the study considers the 11 types of actions discussed in section 3 as a multi-class classification task. For each type of event, the algorithms classify as a "positive" outcome a specific type of event mentioned in the sentence, and "negative" otherwise. This phase applies the same pre-processing steps as in Tasks 1 and 2. To perform the classification, we generated individual binary subsets for each event label, enabling a binary classification for every label. This means that, from the 11 available classes, we created 11 subsets, each one with only two classes: positive and negative. This allows us to evaluate the performance of the learning algorithms with respect to each event type.

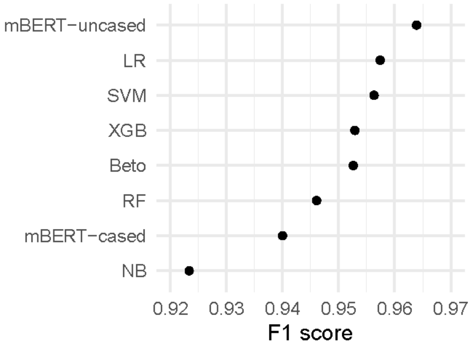


Figure 6: Document binary classification (task 1)

5 Results

5.1 Task 1: Document relevance

Figure 6 presents the results of the binary document classification using a diverse set of ML algorithms and LLM. According to the results, the model with the best performance in classifying relevant and not relevant news articles associated with criminal violence and law enforcement efforts in Spanish is mBERT-uncased with an F1=0.9630.

All other models in the binary document classification task report lower performance than mBERT-uncased. The Logistic Regression (LR) model reports the second-best results with an F1=0.9574, closely followed by the Support Vector Machine (SVM) with an F1=0.9563; the Extreme Gradient Boosting (XGB) model performs with an F1=0.9529; BETO reports an F1=0.9526; the Random Forest (RF) Classifier indicates an F1=0.9461; the cased version of Multilingual BERT (mBERT-cased) model reports an F1=0.94; and finally, the Multinomial Naive Bayes (NB) has the lowest performance with an F1=0.9233.

The result of the mBERT-uncased application in this study considerably outperforms the performance of the Logistic Regression model originally implemented by Osorio and Beltrán (2020) for document classification, which reached an F1=0.949. The high F1 performance of the mBERT-uncased model in this application also stands out with respect to the performance of other binary document classification efforts on similar domains. For example, the highest score of binary document classification of protest data reported in the CASE 2022 joint task reached an F1=0.7496 (Hürriyetoğlu et al., 2022), which is considerably lower than the performance reported in this study. The results of the mBERT-uncased pre-trained language model in Spanish also puts into perspective the results of other studies showing that other models outperform mBERT in Spanish on similar conflict-related domains such as hate speech (Castillo-lópez et al., 2023) and sexism detection (Schütz et al., 2022).

5.2 Task 2: Sentence relevance

As indicated in the experiment setting in section 4, we proceed in an agnostic way with respect to the different ML algorithms and LLM considered in this study and put them all to compete in classifying relevant sentences. The first row of Table 1 reports the results of the different models on the classification of relevant sentences. The performance metric used to assess the models is the average macro-F1 derived from running five iterations of each model. In this way, the results provide evidence of the average performance of each model, rather than arbitrarily picking the top performance from any random seed. The model reporting the highest macro-F1 is marked in bold font to indicate the algorithm that has the best performance in each classification category.

The results of Table 1 show that BETO has the best performance for sentence relevance classification with an F1=0.8588. The model with the second best performance is mBERT-uncased (F1=0.8553), followed by mBERT-cased (F1=0.8506). All other models have slightly lower performance for classifying the relevance of specific sentences.

In general, BERT-like models stand out for their high performance in identifying relevant sentences related to organized criminal violence and law enforcement efforts from text in Spanish. The excellent performance of this model is consistent with the findings of Hürriyetoğlu et al. (2022) for classifying protest data, which reached a maximum F1=0.8245 in its top-performing model.

5.3 Task 3: Event type

Finally, the rest of the rows in Table 1 show the results of the different ML algorithms and LLM on the multi-class classification of specific event types in relevant sentences. In general, the performance of event classification reflects the expectations of unbalanced annotations discussed in the Training Data section 3. As expected, the models generally perform better for event types that have a large number of annotations, while they tend to show lower performance for rare event types.

The second section of Table 1 reports the results of the different ML and LLM tools for actions initiated by organized criminals. The BETO model

			Traditional ML					LLM		
		Positive						mBERT	mBERT	
	Task	cases	NB	SVM	LR	RF	XGB	uncased	cased	BETO
1	Relevant	6,327	0.8179	0.8443	0.8449	0.8512	0.8368	0.8553	0.8506	**0.8588**
2	Criminal violence vs. criminals	396	0.0000	0.2799	0.0379	0.0805	0.2614	0.6630	0.8412	**0.8550**
	Criminal violence vs. state	1,148	0.0522	0.5836	0.4211	0.3120	0.5874	**0.8423**	0.8170	0.8380
	Criminal violence vs. civilians	2,556	0.6840	0.7397	0.7147	0.6563	0.6959	0.8558	0.8536	**0.8699**
	Drug trafficking or production	402	0.0103	0.3908	0.1333	0.1640	0.3865	0.6817	**0.8440**	0.4994
3	State violence vs. criminals	405	0.0000	0.6483	0.4547	0.2234	0.6815	0.8312	0.8457	**0.8505**
	State violence vs. civilians	30	0.0000	0.2133	0.0000	0.2133	0.2133	0.4992	0.8382	**0.8558**
	State arrest of criminal	1,093	0.0569	0.6255	0.4177	0.3727	0.6246	0.8305	0.8441	**0.8550**
	State seizure of assets	490	0.2355	0.5896	0.3204	0.5289	0.5546	0.7757	0.7450	**0.8451**
	State seizure of guns	859	0.7296	0.8268	0.7519	0.7519	0.8497	**0.9191**	0.8421	0.8567
	State seizure of drugs	1,041	0.6444	0.8086	0.7088	0.8133	0.8131	**0.9259**	0.8367	0.8596
4	Civilian violence vs. criminals	19	0.0000	0.0000	0.0000	0.0000	0.0000	0.4997	**0.8440**	0.4994

Table 1: Relevance and multi-class classification at the sentence level (tasks 2 and 3).

has the best performance for classifying criminal violence against criminals with an F1=0.8550. This performance is remarkable given the small number of positive cases in this category. The mBERT-uncased model reports the top performance for classifying incidents of criminal violence against the state with an F1=0.8423. Given the variety of potential targets among state authorities (e.g. soldiers, marines, police, and state officials), it may be difficult for the algorithm to accurately identify incidents of criminal violence against the state. Yet, the substantial number of annotations in this category likely contributes to the model's good performance. For the category related to criminal violence against civilians, BETO reports the best performance with an F1=0.8699. Given the broad range of crime victims and the different types of violent tactics used by criminal groups, it is difficult to accurately classify this category. Finally, the algorithm achieving the best performance when classifying drug trafficking or production is mBERT-cased with an F1=0.8440. This level of performance is remarkable given the limited observations in this category and the broad variety of narcotics (e.g. cocaine, heroin, fentanyl) that make their classification a challenging task.

The third section of Table 1 reports the performance for events initiated by the state. Despite the limited number of annotations of state violence against criminals and civilians, the best-performing model, BETO, reports an F1=0.8505 and F1=0.8558, respectively. BETO is also the top-performing model for classifying arrests of criminals, with an F1=0.8550. According to the results, the algorithm with the best performance at identifying seizures of assets (e.g. vehicles, real estate) is again BETO with an F1=0.8451. Results show an outstanding performance of mBERT-uncased to identify gun and drug seizures with an F1=0.9191 and F1=.0.9259, respectively.

Finally, the bottom section of Table 1 reports the results of identifying incidents of civilian violence against criminals. Due to the extremely low number of annotations in this event-type category, most models struggle with effectively classifying this type of event. However, mBERT-cased reports a strong performance with an F=0.8440.

In general, the results of Table 1 show that traditional ML algorithms have a sub-optimal performance. In contrast, the family of BERT models consistently reports higher levels of performance. This is consistent with the well-documented high-performance of BERT models in a variety of NLP tasks (Devlin et al., 2019).

6 Conclusions

This study presents an application to classify information related to organized criminal violence from unstructured text written in Spanish using ML and LLM. The results from this study enable researchers and government authorities to track the violent behavior of organized criminal groups in Mexico and assess the effects of law enforcement activities. This research allows for the generation of data on a large scale, in a timely manner, and with an unprecedented degree of granularity and accuracy. By analyzing criminal and state behaviors, the study goes beyond previous efforts exclusively focusing on tracking the territorial presence of criminal groups using rule-based approaches (Osorio and Beltrán, 2020; Signoret et al., 2021; Coscia and Rios, 2012). Tracking the territorial presence of criminal groups only provides information about *who* is present, but does not say anything about *what are they doing*. Thus, this study provides valuable tools to identify behavioral trends of criminal groups and state authorities from news stories with unprecedented accuracy.

The methodological approach in this research focuses on a sequence of classification tasks of increasing levels of detail. Based on a large collection of documents and a robust set of high-quality annotations in the training data, the first task focuses on classifying the relevance of entire news articles related to organized criminal violence and law enforcement. To do so, the experimental setting puts a variety of ML algorithms to compete in the binary classification task. The algorithm reporting the best performance is Multilingual BERT - uncased, with an F1 score of 0.9630. This high level of performance provides a strong indicator of the effectiveness of this ML algorithm.

The second stage evaluates the different algorithms for the identification of relevant sentences. Results show that BETO presents the highest level of performance for the binary sentence classification task with an F1 score of 0.8588.

Finally, the study focuses on classifying the specific types of events of organized criminal violence and law enforcement contained in the data. This application considers 11 different types of events of lethal and non-lethal violence initiated by criminal groups, government authorities, and the civilian population. Given the variations in the distribution of annotations across event categories, results show varying degrees of performance in the multi-class classification of event types. In general, the family of BERT-like models shows a strong performance when classifying different types of organized criminal violence and law enforcement efforts. Results of these NLP tasks report F1 scores ranging from 0.8440 to 0.9259. In particular, BETO consistently presents high performance in many categories.

Beyond the technical performance evaluated in this application, results provide great confidence about the use of NLP tools to accurately extract and classify a broad range of behavioral information related to organized criminal violence from text written in Spanish. These results offer valuable contributions to researchers, security analysts, and government agencies in Spanish-speaking countries in their efforts to understand organized criminal behavior using high-quality data.

6.1 Future Work

A key limitation in this study is the combination of imbalanced training data and the low number of annotations in some event categories that undermine performance. In order to overcome this limitation, future research will explore different data augmentation techniques (Şahin, 2022; Yang et al., 2022). In particular, the Confli-T5 method (Parolin et al., 2022) is a promising one as it specializes in political violence and conflict. However, Confli-T5 was developed in English and requires multi-lingual extensions.

Future research should also consider recent developments in pre-training language models relevant to crime, violence, and politics. Recently, Parolin et al. (2021) proposed the 3M-Transformers (Multilingual, Multi-label, Multitask) method to classify and extract information related to crime and conflict in English, Spanish, and Portuguese. Most importantly, future research should consider using ConfliBERT (Hu et al., 2022) and the ConfliBERT variation in Spanish (Yang et al., 2023), a domain-specific language model specialized in conflict and violence. Independent research has shown that ConfliBERT is the state-of-the-art model for NLP tasks on political violence and conflict (Häffner et al., 2023). Unfortunately, the current ConfliBERT version is only capable of processing text in English.

6.2 Ethical considerations

This study was conducted in compliance with the ACL ethical research guidelines and operated under the supervision of the University of Arizona IRB (protocol 2012326746A001). This project only used secondary data and did not involve human research subjects. Also, as discussed in section 3, the coding protocol took extra measures to avoid further stigmatizing crime victims.

Acknowledgments

This research was possible thanks to the support of the Technology and Research Initiative Fund of the University of Arizona, the National Science Foundation [SES-2116406, OAC-2311142], and the U.S. Department of Defense [W911NF-20-1-0303]. Thanks to Mariana Marín, Rubén Oliva, and Íñigo Casares for their research support. The replication files of this study are available at https://github.com/javierosorio/classifying_criminal_violence_mx

References

Elizabeth Boschee, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz, and Michael Ward. 2016. ICEWS Coded Event Data. Publication Title: Harvard Dataverse.

Leo Breiman. 2001. Random Forests. *Machine Learning*, 45(1):5–32.

Tommaso Caselli, Osman Mutlu, Angelo Basile, and Ali Hürriyetoğlu. 2021. PROTEST-ER: Retraining BERT for Protest Event Extraction. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 12–19, Online. Association for Computational Linguistics.

Galo Castillo-lópez, Arij Riabi, and Djamé Seddah. 2023. Analyzing Zero-Shot transfer Scenarios across Spanish variants for Hate Speech Detection. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 1–13, Dubrovnik, Croatia. Association for Computational Linguistics.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ArXiv:1603.02754 [cs].

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Michelle Coscia and Viridiana Rios. 2012. Knowing Where and How Criminal Organizations Operate Using Web Content. In *CIKM'12. ACM international conference on Information and knowledge management*, volume October, Maui, HI.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Aaron Erlich, Stefano G. Dantas, Benjamin E. Bagozzi, Daniel Berliner, and Brian Palmer-Rubin. 2022. Multi-Label Prediction for Political Text-as-Data. *Political Analysis*, 30(4):463–480. Publisher: Cambridge University Press.

Justin Grimmer and Brandon M. Stewart. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297.

Andy Halterman, Philip A Schrodt, Andreas Beger, Benjamin E. Bagozzi, and Grace Scarborough. 2023a. Creating Custom Event Data Without Dictionaries: A Bag-of-Tricks.

Andy Halterman, Philip A Schrodt, Andreas Beger, Benjamin E. Bagozzi, and Grace Scarborough. 2023b. PLOVER and POLECAT: A New Political Event Ontology and Dataset.

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, HLT '09, pages 27–35, USA. Association for Computational Linguistics.

Yibo Hu, MohammadSaleh Hosseini, Erick Skorupa Parolin, Javier Osorio, Latifur Khan, Patrick Brandt, and Vito D'Orazio. 2022. ConfliBERT: A Pre-trained Language Model for Political Conflict and Violence. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5469–5482.

Sonja Häffner, Martin Hofer, Maximilian Nagl, and Julian Walterskirchen. 2023. Introducing an Interpretable Deep Learning Approach to Domain-Specific Dictionary Creation: A Use Case for Conflict Prediction. *Political Analysis*, pages 1–19. Publisher: Cambridge University Press.

Ali Hürriyetoğlu, Osman Mutlu, Fırat Duruşan, Onur Uca, Alaeddin Gürel, Benjamin J. Radford, Yaoyao Dai, Hansi Hettiarachchi, Niklas Stoehr, Tadashi Nomoto, Milena Slavcheva, Francielle Vargas, Aaqib Javid, Fatih Beyhan, and Erdem Yörük. 2022. Extended Multilingual Protest News Detection - Shared Task 1, CASE 2021 and 2022. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 223–228, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Canete José, Chaperon Gabriel, Fuentes Rodrigo, and Pérez Jorge. 2020. Spanish pre-trained BERT model and evaluation data. *PML4DC at ICLR*, 2020.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text classification algorithms: A survey. *Information*, 10(4):150.

Maria Krommyda, Anastasios Rigos, Kostas Bouklas, and Angelos Amditis. 2021. An Experimental Analysis of Data Annotation Methodologies for Emotion Detection in Short Text Posted on Social Media. *Informatics*, 8(1):19. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

Will Lowe and Kenneth Benoit. 2013. Validating estimates of latent traits from textual data using human judgment as a benchmark. *Political analysis*, 21(3):298–313.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3):1–40.

Claire Moon and Javier Treviño-Rangel. 2020. "Involved in something (involucrado en algo)": Denial and stigmatization in Mexico's "war on drugs". *The British Journal of Sociology*, 71(4):722–740. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-4446.12761.

Petr Motlicek. 2023. ROXANNE - Real time network, text, and speaker analytics for combating organized crime.

Javier Osorio. 2015. The Contagion of Drug Violence: Spatiotemporal Dynamics of the Mexican War on Drugs. *Journal of Conflict Resolution*, 59(8):1403–1432.

Javier Osorio and Alejandro Beltrán. 2020. Enhancing the Detection of Criminal Organizations in Mexico using ML and NLP. *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. Glasgow, Scottland.

Javier Osorio, Mohamed Mohamed, Viveca Pavon, and Brewer-Osorio Susan. 2019. Mapping Violent Presence of Armed Actors. *Advances in Cartography in GIScience of the International Cartographic Association*, pages 1–16.

Javier Osorio and Alejandro Reyes. 2017. Supervised Event Coding From Text Written in Spanish: Introducing Eventus ID. *Social Science Computer Review*, 35(3):406–416.

Erick Skorupa Parolin, Yibo Hu, Latifur Khan, Patrick T. Brandt, Javier Osorio, and Vito D'Orazio. 2022. Confli-T5: An AutoPrompt Pipeline for Conflict Related Text Augmentation. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 1906–1913.

Erick Skorupa Parolin, Latifur Khan, Javier Osorio, Patrick Brandt, Vito D'Orazio, and Jennifer Holmes. 2021. 3M-Transformers for Event Coding on Organized Crime Domain. *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. Publisher: IEEE.

Pedro L Rodriguez and Arthur Spirling. 2022. Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics*, 84(1):101–115.

Elena Rudkowsky, Martin Haselmayer, Matthias Wastian, Marcelo Jenny, Štefan Emrich, and Michael Sedlmair. 2018. More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, 12(2-3):140–157.

Philip A. Schrodt, Brandon Stewart, Jennifer Lautenschlager, Andrew Shilliday, David Van Brackel, and Will Lowe. 2010. Automated Production of High-Volume, Near-Real-Time Political Event Data. Technical report. Publication Title: Event (London).

Philip A. Schrodt and David Van Brackel. 2013. Automated Coding of Political Event Data. In Devika Subramanian, editor, *Handbook of Computational Approaches to Counterterrorism*, pages 23–50. Springer, New York.

Mina Schütz, Jaqueline Boeck, Daria Liakhovets, Djordje Slijepčević, Armin Kirchknopf, Manuel Hecht, Johannes Bogensperger, Sven Schlarb, Alexander Schindler, and Matthias Zeppelzauer. 2022. Automatic Sexism Detection with Multilingual Transformer Models. ArXiv:2106.04908 [cs].

Patrick Signoret, Marco Alcocer, Cecilia Farfan-Mendez, and Fernanda Sobrino. 2021. Mapping Criminal Organizations.

Zhanna Terechshenko, Fridolin Linder, Vishakh Padmakumar, Michael Liu, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. 2020. A comparison of methods in political science text classification: Transfer learning language models for politics. *Available at SSRN 3724644*.

Muthuraman Thangaraj and Muthusamy Sivakami. 2018. Text classification techniques: A literature review. *Interdisciplinary journal of information, knowledge, and management*, 13:117.

Michael Ward, Andreas Beger, Josh Cutler, Matthew Dickenson, Cassy Dorff, and Ben Radford. 2013. Comparing GDELT and ICEWS Event Data.

Guanqun Yang, Mirazul Haque, Qiaochu Song, Wei Yang, and Xueqing Liu. 2022. TestAug: A Framework for Augmenting Capability-based NLP Tests. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3480–3495, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Wooseong Yang, Sultan Alsarra, Lujay Abdeljaber, Niamat Zawad, Zeinab Delaram, Javier Osorio, Latifur Khan, Patrick T. Brandt, and Vito D'Orazio. 2023. ConfliBERT-Spanish: A Pre-trained Spanish Language Model for Political Conflict and Violence. In *Proceedings of the 2023 Recent Advances in Natural Language Processing conference*, Varna, Bulgaria.

Gözde Gül Şahin. 2022. To Augment or Not to Augment? A Comparative Study on Text Augmentation Techniques for Low-Resource NLP. *Computational Linguistics*, 48(1):5–42.