Classifying Narcotrafficking Spatial Event Documents using Transformers

Morteza Karimzadeh[1][0000-0002-6498-1763], Huilin Han[1], Beth Tellman[2][0000-0003-3026-6435], Erik Nielsen[3]

[1] University of Colorado Boulder, Boulder, CO, 80302, USA, karimzadeh@colorado.edu
[2] University of Arizona, Tucson, AZ 85721, USA
[3] Northern Arizona University, Flagstaff, AZ 86011, USA

Abstract. The low signal-to-noise ratio of information in unstructured textual documents remains a challenge in geo-text analyses. While information extraction approaches can be used to identify places, times and people mentioned in text, many of the identified entities may not be related to the analytical scenario, and may mislead spatial models without pre-filtering. On the other hand, detecting relevant spatiotemporal events and their attributes is of high interest to geovisual or computational solutions leveraging textual data. In this paper, we present a classification approach for identifying documents describing the time, locations and attributes of such spatiotemporal events (in Spanish), namely of drug trafficking activities in Honduras. We fine-tune a Spanish-specific and a Multilingual BERT (Bidirectional Encoder Representations from Transformers) model for this task with our limited amounts of training data. Our results indicate high performance of this approach, and the ability of the models to identify and filter documents including spatiotemporal events. The results are noteworthy since all documents in the dataset are related to narcotrafficking events (regardless of class membership), but not all describe a spatiotemporal event, yet the models exhibit high performance in detecting the ones describing the event.

Keywords: geo-text, spatiotemporal event detection, BERT, narcotrafficking

DOI: https://doi.org/10.25436/E2B88Q

1 Introduction

With recent advances in Natural Language Processing (NLP), automated text analytical pipelines are increasingly used in fields leveraging geo-text analysis. However, the abundance of textual documents, while providing an opportunity for generating data with higher spatiotemporal coverage, presents a challenge in most analytical scenarios: low signal-to-noise ratio of information. The challenge of high noise also manifests itself when attempting to retrieve documents containing spatiotemporal event types of interest. Such documents constitute a very small percentage of a corpus, even after pre-filtering with queries. On the other hand, Named Entity Recognition (NER) methods can extract entities

such as times or locations, but the high majority of those entities may still be unrelated to the spatiotemporal event type of interest. More advanced strategies are required to efficiently filter and use unstructured data in analytical scenarios.

Machine Learning (ML) can be used to reduce the document set to the ones containing spatiotemporal events of interest, provided that task-specific models or sufficient training examples exist. However, ML models are trained and benchmarked on datasets with general topics that are separable in semantic space, and not for separating documents that contain spatiotemporal event descriptions (and are otherwise close to each other in the semantic space). In this study, we present preliminary work on a classification approach using fine-tuning a pre-trained transformer model to detect spatiotemporal narcotrafficking events in Central American Spanish news. A document with a positive label in our study is one that contains information on the time, place and narco-attributes of a trafficking activity. Detecting these events is of high interest to stakeholders and researchers.

Data on illicit transactions are often incomplete, fragmented, or unreliable, whether they are based on official statistics or data leaks (e.g. The Panama Papers) [1]. Beyond official statistics, which are known to underreport drug trafficking activity, media reports are one of the only sources of data that document illicit activities at national- and fine-spatiotemporal scale, including on years-worth of data missing in official sources [2], [3].

However, current methods leveraging news media for illicit-activity quantification largely rely on manual searching and annotation (i.e., coding) of articles. The high labor cost of over one hour per one positively identified event (e.g., a spatiotemporal trafficking event) yields relatively small sample sizes (eg. <1000 observations) [3]. This small sample size both limits the scope of analysis, and also, poses potential challenges to automation using ML. Automation can substantially increase the spatiotemporal coverage and sample size of illicit activity databases across several news media sources. These datasets can then be used to build quantitative causal inference models to evaluate the effectiveness of narcotrafficking interdiction policies, understand (and predict) how narcotrafficking transforms socio-environmental systems, and how to mitigate these negative impacts.

2 Background and Related Work

In a recent study, researchers manually coded narcotrafficking events documented in news media across Central America [3], and found 826 unique events for Honduras at the department (sub-national) spatial scale. Official drug trafficking data, from the Consolidated Counterdrug Database (by US Army SOUTHCOM) covers only 4 departments in Honduras for limited years, yielding 30 department-year observations (McSweeney et al 2020). News media manually coded yielded 118 department-year observations, covering 13 departments (Fig. 1). Where official data were available, it was spatially and temporally correlated with news media counts of narcotrafficking events at both the country (R^2=0.70) and department scale (r ranging from 0.48-0.74 in Honduras). However, the process for detecting spatiotemporal narco-trafficking events in news media is laborious, with approximately one hour per each positively identified document in the above-cited study. Filtering articles using a set of narco-related keywords resulted in a large number of articles, most of which did not describe spatiotemporal specifics of a trafficking activity, although most were generally related to the topic of narco-trafficking.

While spatiotemporal event detection can be mapped to several ML task types, in this article, we focus on classification for this purpose (per request by domain researchers) to identify the small minority of narco-events among narco-related documents. After this identification, it is conceivable that other spatial methods or question/answering can be applied to glean structured knowledge on the specific events.

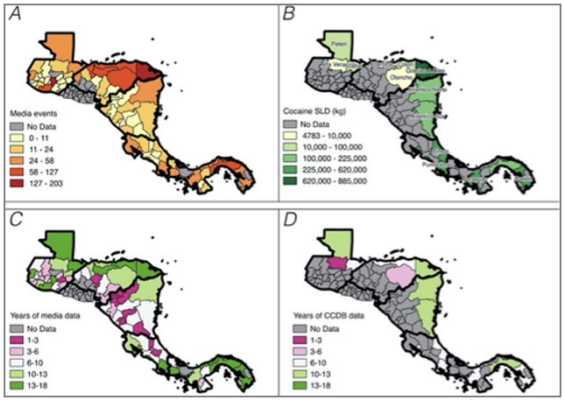

Fig. 1. Media and Consolidated Counterdrug Database (CCDB) data coverage over Central America. A) Total media events per department from 2000-2017, B) Total kilos of cocaine seized, lost, or delivered from 2000-2014, C) Completeness measured in years of media observations present per department, D) Completeness measured in years of CCDB data.

Shortage of labeled training data remains a challenge in some domains, and thus, the last decade of core NLP research has placed much focus on creating pre-trained models. Using this approach, for a task with a limited number of domain labeled examples (as is the case in our study), classification is usually performed by fine-tuning a pre-trained model. The pre-trained models are trained on large datasets (i.e., millions of documents), and made available to the community. The model parameters from training on such huge datasets encode the syntactic and/or semantic relations and dependencies between different words, phrases and sentences. These "frozen" model parameters can then be used by domain researchers and practitioners, where usually fewer number of parameters are unfrozen and re-learned for the specific task at hand [4]. Updating weights/parameters from pre-learned ones (instead of random initial weights) allows for high performance with a small training dataset [5].

One such pre-trained model that has achieved state-of-the-art performance in many NLP tasks is the Bidirectional Encoder Representations from Transformers (BERT) [6]. BERT is trained on two unsupervised tasks. In the first task a percentage of tokens (words) are masked and the model predicts the masked tokens. The second task is next sentence prediction (NSP), where the model is given two sentences A and B, and the model predicts if B follows A in the text. This helps the model encode the syntactic and semantic relations in text. Other language-specific BERT models have been pre-trained, including BETO, a Spanish BERT model that has shown performance above or close to the performance of base multilingual BERT model [7]. BERT can be fine-tuned using various strategies, however, updating the penultimate layer achieves high-performance by leveraging the knowledge learned by BERT [4].

3 Data and Models

We use the manually-labeled dataset of news media reports of drug trafficking events in Honduras generated in [3]. The dataset includes media reports from national newspapers, including Proceso, La Prensa, La Tribuna, El Heraldo, and the book Bribes, Bullets, and Intimidation. Articles published between 2000 and 2017 were retrieved using 33 keywords, for instance, 'narco pista'. Human coders identified articles that describe the place, time, and other attributes of a drug trafficking activity. Since the coding was not performed with the purpose of training ML, the articles that did not have information on time/place/attribute of a trafficking event, were discarded. Thus, it only included "positive" examples.

We retrieved negative examples from La Prensa Honduras by querying the same list of keywords combined with a year between 2007 and 2017, because La Prensa only has online data starting from 2007. We stratified the results across years to ensure diversity of negative examples. We added articles from the first five pages of the search results of each keyword-year combination to the negative set, excluding duplicates and articles present in the positive dataset. The positive and negative examples were scraped and cleaned in the same format, and quality controlled to ensure structural similarity between positive and negative examples. We also restricted positive examples to the ones published in La Prensa to control for potential stylistic variation across newspapers. The resulting dataset consists of 371 positive and 4523 negative examples, each a document with the title and body of news reports.

We fine-tuned two different BERT-structured models for classifying pos/neg examples, by further updating all the pre-trained BERT model parameters on your training dataset, using the SimpleTransformers python library:

3.1 BETO. BETO is pre-trained on a large Spanish corpus consisting of Wikipedia, UN journals, news stories and more, totaling ~3 billion words [7]. BETO is of a similar size to the original BERT-base model, containing 12 self-attention layers and 16 attention heads each, with 768 as hidden size, totaling 110M parameters. BETO achieved state-of-the-art on multiple Spanish NLP benchmark tasks including named entity recognition (NER-C) and document classification (MLDoc).

3.2 Multi-Lingual BERT. Multilingual BERT is a BERT-based model trained on a multilingual corpus from Wikipedia, the newest release supporting 104 languages with comparatively large Wikipedias [6]. Multilingual BERT also has the same architecture and size as the BERT-base model. Multilingual BERT's performance for high-resource languages such as English and Chinese is usually slightly worse than a specialized one-language model. In a comparison between BETO and Multilingual BERT's performance on Spanish NLP benchmarks, BETO generally outperformed Multilingual, albeit marginally in some cases [7].

4 Experiments and Evaluation Results

The narcotrafficking analysts prioritize recall over precision, i.e., fewer false negatives even at the cost of more false positives. Therefore, after some experimentation with different weights, we set the penalty of misclassifying a positive example 60

times higher than a negative example using a weighted binary cross-entropy loss function. Since the negative examples were discarded in the original study, we did not have statistics on the representative ratio of negative to positive (neg2pos) examples. Therefore, we also experimented with a combination of different neg2pos ratios. In Table 1, a 1:1 neg2pos means 371 positive examples and 371 negative examples, while in 1:4, those numbers corresponded to 371 and 1484, respectively. We used 90%-10% as training-testing split. We fine-tuned BETO and Multilingual models separately, with a batch size of 32, and learning rate of 3×10^{-5} for 10 epochs.

Table 1. Precision (precision), recall (re), F_1-score, and Area Under the ROC Curve (AUC) for different models and negative 2 positive ratios.

Model	neg2pos	PR	Recall	F1	AUC
BETO	1	**0.86**	0.90	**0.88**	0.91
BETO	4	0.77	0.70	0.74	0.94
BETO	8	0.64	0.62	0.63	0.93
Multilingual	1	0.81	**0.97**	**0.88**	0.92
Multilingual	4	0.75	0.68	0.71	0.93
Multilingual	8	0.49	0.53	0.54	0.85

Both models achieve the highest F_1 of 0.88, but Multilingual seems to slightly outperform BETO with higher recall and AUC. The 60:1 weight of positive examples made oversampling of negative examples unnecessary, as can be seen by the best results achieved for the 1 neg2pos ratio. The variation in metrics can be attributed to the small size of the testing set. However, the results are promising. The model is at least 90% successful in flagging documents that contain spatial, temporal and attribute information on a trafficking activity, even though all pos/neg articles (positive or negative) are in fact "related to" narcoactivities.

Closer qualitative examination of results indicated that the performance might even be better, as several of the false positives were in fact positive examples, i.e., actual narcotrafficking events but absent from the original dataset. As noted earlier, we had to re-collect negative examples separately from the original dataset that contained only positive examples. The results indicate that the models are finding dozens of new narcotrafficking events. For example, in BETO results, reading the apparent false positive sample of "Barco traía 700 kilos de cocaína en láminas (article #491)" clearly indicates a trafficking event of 700 kilos of cocaine crossing the Bahia of Honduras. Another example "Honduras: Narcoavioneta cae en pista de Utila" describes an airplane full of cocaine crashing in Utila, Honduras, but was included in the negative examples and thus counted in the false positive statistics. For the Multilingual model, the example "Decomisan 214 kilos de cocaína en el este de Guatemala" describes authorities capturing 214 kilos of cocaine, and therefore, is a real narcotrafficking event, but then again, included in the false positive statistics due to the example coming from the negative set. All articles inspected identified as true positive and true negative were deemed correctly identified by the model.

5 Conclusion

Our initial results indicate that fine-tuning a BERT model allows for accurate and efficient identification of documents containing spatiotemporal events. In our case, this was notable since all documents (positive or negative) were related to the topic of narcotrafficking, and our training dataset was small; nevertheless, resulting in satisfactory performance for identifying spatiotemporal events and discovering new ones missed in manual searches in previous work.

Future work includes assessing if accurate spatially-explicit information extraction is possible, namely, the extraction of time and location of the narcotrafficking event at fine scale. Our results indicate the high promise of such approach, as the signals associated with the spatiotemporal descriptions of the narco-events seem to have been detected by BERT features, resulting in the high performance of classification. Therefore, it is likely that the entity extraction results would be favorable, provided that first narco-articles are filtered out of the massive numbers of news docs using our approach. Quantitative evaluation of more advanced models requires the annotation of entities within documents, which we will undertake in future research. The findings have implications for geo-text analyses benefiting from the detection of spatiotemporal events in text. Another noteworthy observation is the high performance of multi-lingual deep models, similar to the language-specific ones. While Spanish is a high-resource language, low-resource languages' corpora are up-sampled in multi-lingual BERT, potentially providing an opportunity for leveraging text in other languages.

Acknowledgements

This work was partially funded by UROP at the University of Colorado. We thank the students who coded the media at Northern Arizona University: Leah Manak, Hannah Russell, Alana Weber, and Rafael Ramirez, funded by Dr. Erik Nielsen.

References

1. A. Rege, "Not biting the dust: using a tripartite model of organized crime to examine India's Sand Mafia," Int. J. Comp. Appl. Crim. Justice, 2016.
2. R. Hudson, "Thinking through the relationships between legal and illegal activities and economies: Spaces, flows and pathways," J. Econ. Geogr., 2014.
3. B. Tellman et al., "Illicit Drivers of Land Use Change: Narcotrafficking and Forest Loss in Central America," Glob. Environ. Chang., 2020..
4. C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," in Lecture Notes in Computer Science. 2019.
5. L. S. Snyder, Y. Lin, M. Karimzadeh, D. Goldwasser, and D. S. Ebert, "Interactive Learning for Identifying Relevant Tweets to Support Real-time Situational Awareness," IEEE Trans. Vis. Comput. Graph., p. 1, 2019, doi: 10.1109/TVCG.2019.2934614.
6. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 2019.
7. J. Cañete, G. Chaperon, R. Fuentes, and J. Pérez, "Spanish Pre-Trained BERT Model and Evaluation Data," to Appear PML4DC ICLR 2020, 2020.