ConfliBERT-Spanish: A Pre-trained Spanish Language Model for Political Conflict and Violence

Wooseong Yang
Dept. of Computer Science
University of Texas at Dallas
Richardson, USA
wooseong.yang@utdallas.edu

Sultan Alsarra
Dept. of Computer Science
University of Texas at Dallas
Richardson, USA
sultan.alsarra@utdalls.edu

Luay Abdeljaber
Dept. of Computer Science
University of Texas at Dallas
Richardson, USA
luay.abdeljaber@utdallas.edu

Niamat Zawad
Dept. of Computer Science
University of Texas at Dallas
Richardson, USA
niamat.zawad@utdallas.edu

Zeinab Delaram
Dept. of Computer Science
University of Texas at Dallas
Richardson, USA
zeinab.delaram@utdallas.edu

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

Latifur Khan
Dept. of Computer Science
University of Texas at Dallas
Richardson, USA
lkhan@utdallas.edu

Patrick T. Brandt
School of Economic, Political, and Policy Sciences
University of Texas at Dallas
Richardson, USA
pbrandt@utdallas.edu

Vito D'Orazio
School of Mathematical and Data Sciences
West Virginia University
Morgantown, USA
vito.dorazio@mail.wvu.edu

2023 7th IEEE Congress on Information Science and Technology (CiSt) | 978-1-6654-6133-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/CiSt56084.2023.10409883

***Abstract*—This article introduces ConfliBERT-Spanish, a pre-trained language model specialized in political conflict and violence for text written in the Spanish language. Our methodology relies on a large corpus specialized in politics and violence to extend the capacity of pre-trained models capable of processing text in Spanish. We assess the performance of ConfliBERT-Spanish in comparison to Multilingual BERT and BETO baselines for binary classification, multi-label classification, and named entity recognition. Results show that ConfliBERT-Spanish consistently outperforms baseline models across all tasks. These results show that our domain-specific language-specific cyberinfrastructure can greatly enhance the performance of NLP models for Latin American conflict analysis. This methodological advancement opens vast opportunities to help researchers and practitioners in the security sector to effectively analyze large amounts of information with high degrees of accuracy, thus better equipping them to meet the dynamic and complex security challenges affecting the region.**

***Index Terms*—NLP, Deep Learning, BERT, Machine Learning, Spanish, Politics, Conflict, Violence**

I. Introduction

According to the United Nations, Latin America is the most violent region of the world [1]. Ravaged by political and criminal violence, political crises, drug trafficking, social unrest, and mass migration, researchers and authorities need highly accurate cyberinfrastructure to process massive amounts of information to meet these security challenges in an effective and timely manner. An effective security-oriented cyberinfrastructure in Spanish could potentially help to save lives and reduce the deleterious impacts of violence in Latin America. Based on early Natural Language Processing (NLP) efforts using rule-based approaches in Spanish language [2], [3], researchers have been developing more advanced Machine Learning (ML) tools to study political violence [4], [5], conflict forecasting [6], organized crime [7], and social unrest [8]. Unfortunately, many of these developments lack the right domain specificity in Spanish to find effective solutions to complex NLP challenges.

Recent pre-trained language models like BERT [9] revolutionized our capacity to tackle a variety of NLP tasks such as text classification, named entity recognition, information extraction, sentiment analysis, and language synthesis. Unfortunately, many of these tools exclusively work in the English language and their application to Spanish is limited despite having about 500 Million speakers worldwide [10]. Part of the challenge of Spanish NLP rests on the pronoun-drop character of Spanish. In contrast to non-pronoun-drop languages such as English, Spanish often omits the pronoun to enable a richer inflection for the person, number, gender, and tense into the verb conjugation. This pronoun-drop inflection makes Spanish NLP analysis particularly challenging.

To address this gap, this research introduces ConfliBERT-Spanish, a pre-trained language model specialized in political conflict and violence for text written in the Spanish language. Based on the initial developments of ConfliBERT-English [11] and ConfliBERT-Arabic [12], our methodological approach pushes the technological frontier of Multilingual BERT [9] and BETO [13], to generate a domain-specific language-specific language model capable of tackling complex NLP task relevant to political conflict and violence in Spanish language.

To develop ConfliBERT-Spanish, we gathered and curated a large corpus specialized on politics and violence from multiple sources in Spanish. Our methodology then used this domain-specific corpus to pre-train ConfliBERT-Spanish using a continual approach based on Multilingual BERT and BETO, the two most prevalent generic models in Spanish. By doing so, we take advantage of the vast learning of those pre-existing models and develop an enhanced capacity to process challenging tasks in political violence in Spanish.

We compare ConfliBERT-Spanish against the Multilingual BERT and BETO baselines in a variety of domain-relevant tasks that include binary classification, multi-label classification, and named entity recognition. Results consistently show that ConfliBERT-Spanish outperforms generic Multilingual BERT and BETO across all tasks. These results are consistent with studies showing performance improvements by domain-specific pre-trained language models in other fields [14]–[19].

The article proceeds with the following structure. First, we review relevant research pertinent to our work. Then, we describe the methodology we implemented to gather our domain-specific corpus, pre-train our ConfliBERT-Spanish model, and set up the evaluation tasks. The following section presents the results of the evaluation. Finally, we conclude by summarizing our contributions and presenting future research directions.

II. Related Work

Recent pre-trained language models have revolutionized NLP tools. We discuss how four of the most well-known models served as building blocks for ConfliBERT Spanish.

A. BERT

The Bidirectional Encoder Representations from Transformers (BERT) language model [20] is based on a sizable corpus of generic text and uses an unsupervised learning technique that to predict a word randomly absent in a phrase. Due to its excellent performance, BERT quickly became the industry standard for a variety of NLP tasks including sentiment analysis, question-answering, classification, and translation. To improve BERT´s performance for certain tasks, a task-specific layer is placed on top of the pre-trained model, and the entire architecture is trained on a labeled dataset. This method has proven to produce cutting-edge outcomes in a number of NLP tasks. However, BERT's emphasis on the English language limits its potential in non-English NLP analysis.

B. Multilingual BERT

To address BERT's monolingual limitations, researchers developed Multilingual BERT (mBERT) [9] using a 110k common WordPiece vocabulary collection that covers 104 languages for tokenization. Just like the English BERT, mBERT uses lowercase and accent removal, punctuation splitting, and whitespace tokenization. However, intricate morphological systems in certain languages make it difficult for tokenization to be BERT-compatible, which causes redundancy [21]. Subsequent evaluations show that mBERT performs better than the original BERT in a number of languages [22].

C. BETO

BETO is a pre-trained language model specifically designed for Spanish language understanding [13]. BETO is trained on a large Spanish corpus using the same methodology as BERT but uses a different tokenization strategy, taking into account the unique characteristics of the Spanish language. Like BERT, BETO shows promising results in various NLP tasks in Spanish. Both mBERT and BETO helped to advance NLP for Spanish applications but, just like BERT, all these models are developed using non-domain-specific text focused on political conflict and violence.

D. ConfliBERT

Research shows that generic language models tend to underperform in domain-specific NLP tasks involving specialized language. For this reason, scholars have been developing domain-specific pre-trained language models for biology and medicine [14], [15], law [16], science [17], and even the Dark web [19]. All these models tend to outperform generic BERT in domain-specific tasks.

Following the trend of domain-specific models, Hu et al. [11] developed ConfliBERT, a pre-trained language model specialized in political conflict violence. ConfliBERT uses a multitask learning technique on a large domain-specific corpus to carry out numerous related tasks at once. ConfliBERT shows excellent performance on a number of challenges involving the detection of political violence. Similar to other developments, ConfliBERT uses a task-specific layer on top of the pre-trained model for fine-tuning and trains the complete architecture on a labeled dataset for downstream tasks.

Every model has its strengths and limitations. Although BERT excels in several NLP tasks, it has a narrow concentration on English language and its generic training faces limitations in domain-specific jobs. In contrast, mBERT and BETO advance the language frontier by focusing on Spanish, but their training is not domain-specific. Despite its contributions to conflict research, ConfliBERT is also monolingual and it faces limitations in NLP tasks beyond English.

To develop a model specialized in the Spanish language and the field of political conflict and violence, our work tries to build on the advantages of earlier language models. By integrating the features of BERT, BETO, and ConfliBERT, we aim at developing a Spanish-specific and conflict-specific language model capable of accurately identifying and analyzing instances of political violence from text in Spanish.

III. Approach

To develop ConfliBERT-Spanish, we gathered a domain-specific corpus, pre-trained our language model, and evaluated it on several downstream tasks (see Figure 1). Our methodology begins by constructing a Spanish corpus specialized in political conflict and violence. Since publicly accessible Spanish datasets in this domain are scarce, we collected our own pre-training data. By gathering Spanish conflict texts from multiple sources, we developed a dataset that caters specifically to our needs, resulting in improved performance

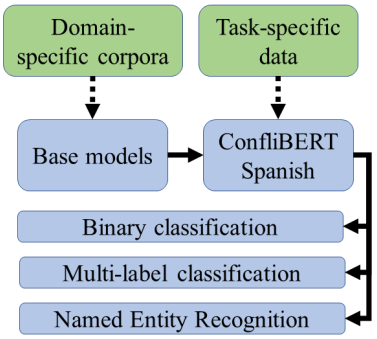

Fig. 1: ConfliBERT-Spanish workflow

on political and violence tasks. After building the corpus, we developed a Spanish-conflict-specific language model using the BERT-based structure, which has shown successful performance [9], [11], [13]–[17]. After pre-training ConfliBERT Spanish, our methodology evaluates the model on various downstream tasks related to political conflict and violence. The subsections below detail these steps.

A. Corpora Building

To guide the data gathering efforts, we use the CAMEO ontology [23] that defines conflict as the dynamics of material/verbal-conflict/cooperation between political entities. Our corpora integrate 11.7 GB of conflict-related texts from various sources. First, we crawled Spanish news websites, only focusing on relevant categories. Also, we crawled the websites of Non-Government Organizations (NGOs) specializing on violence, crime, human rights, and politics in Spanish-speaking countries. Lastly, we used Spanish documents from the United Nations using MultiUN [24], and the European Union's Directorate-General for Translation (DGT) [25].

1) News websites: We sourced political conflict and violence text from 123 news websites in 18 Spanish-speaking countries. We crawled the "politics" and "international" sections that contain a large amount of conflict-related data. Although we focused only on relevant news categories, we could still observe irrelevant texts in the corpus. For example, some articles cover the Olympic games, which often use bellicose language similar to that used in conflict stories. To improve the corpus validity by eliminating non-conflict stories, we filtered the text using relevant and irrelevant keywords. The relevant list came from the CAMEO dictionaries [23], [26] and our own team of conflict scholars created the irrelevant list. The total size of the Spanish newspaper corpus is 7.8 GB.

2) NGO websites: We crawled conflict-relevant reports and documents from NGO websites written in Spanish. We selected NGO websites that deal with political conflict, violence, crime, and human rights, which are highly relevant topics to our domain. We collected text from 97 NGOs in 8 countries and it reached approximately 1.1 GB in size.

3) Public dataset: We used public Spanish text from MultiUN [24] and DGT [25]. MultiUN is a Multilingual corpus of United Nations documents translated into multiple languages. DGT dataset is a Multilingual corpus by the European Union's Directorate-General for Translation. The datasets mostly include political and conflict-related content reflecting the activities of the organizations that developed them. The Spanish text from these databases is approximately 2.8 GB.

TABLE I: Statistics related to the conflict-specific dataset

Parameters	Count
Number of Articles	8,275,941
Words (Tokens)	2,070,287,605
Number of Sentences	52,485,055
Average Number of Words Per Sentence	39.45
Overall Token Frequency	6,238,240

Table I describes some of the characteristics of our aggregated domain-specific corpora after filtering.

B. Pre-training

There are two strategies for a pre-training model on a domain-specific corpus: from scratch (scr) and continual (cont). The scr strategy refers to pre-training model on the corpus, starting from zero base, without using the information learned in pre-existing models. On the other hand, the cont strategy takes advantage of the learning contained in the existing model and continues to pre-train it using an additional domain-specific corpus.

We employed a cont method to fine-tune BERT-based models to the political conflict and violence domain. For the method, we used mBERT and BETO as the base models. This involves initializing the model's vocabulary and checkpoints and training them for additional steps using our domain-specific corpus specialized in political conflict and violence. Since mBERT and BETO have already been pre-trained on a generic domain for Spanish, the cont method requires fewer steps than training a new model from scratch.

C. Evaluation

Most works on pre-trained language models use comprehensive benchmarks in the general NLP [27]–[30]. However, these generic and English-based benchmarks are not appropriate for evaluating our Spanish-domain-specific work. As ConfliBERT-Spanish is originally designed to deal with Spanish text in political conflict and violence, it requires benchmarks valid for the Spanish language and in the conflict-related domain. To address the evaluation challenge of our model, we collected a diverse set of datasets related to political conflict and violence, to conduct the evaluation on various downstream tasks. Next, we describe the datasets used for each downstream task.

1) Binary Classification (BC): Binary Classification is a basic but essential task for conflict scholars to classify domain-relevant documents from large-scale news articles. For the BC task, we built two datasets. For the **Huffingtonpost Spanish** dataset, we crawled text from Huffingtonpost Spanish website and gathered news articles from the politics and international categories, and labeled them as politically relevant articles.

In contrast, we labeled articles from the sports and economy categories as not relevant. For the **Protest** dataset, we collected articles from Spanish Gigaword [31], a large collection of news articles in Spanish. To annotate the data we used the Quad-Class classification from CAMEO to categorize events as (1) material conflict, (2) verbal conflict, (3) material cooperation, and (4) verbal cooperation. For the BC task, we consider both material conflict and verbal conflict as relevant categories, and material cooperation and verbal cooperation as not relevant.

2) Multi-label Classification (MLC): Multi-label Classification can be a useful task for conflict researchers as they are often interested in analyzing different types of conflicts. For the MLC task, we collected two datasets. We first used the **InsightCrime** dataset [32] that consists of news articles reporting organized crime activity in Spanish. We selected the four most common labels out of an original list of 17 labels. Therefore, the MLC task on the InsightCrime dataset classifies stories as "Law Enforcement", "Drug Trafficking", "Homicides", and "Corruption". We also used the **Protest** dataset for MLC classification of the four QuadClass categories: "material conflict", "material cooperation", "verbal conflict", and "verbal cooperation".

3) Named Entity Recognition (NER): As the goal of Named Entity Recognition task is to recognize and classify named entities in a given text, it can definitely be helpful for researchers to detect conflict events and political actors. For NER task, we used **Mx-News** dataset [33] based on political news in Spanish and it considers seventeen classes for entities, including people names (PER), organizations (ORG), dates (DAT), title (TIT), toponyms (GPE), political party (PEX), time expressions (TIM), facility names (FAC), event names (EVT), addresses (ADD), monetary amounts (MNY), laws (DOC), product names (PRO), percentage expressions (PRC), racial characteristics (DEM), age (AGE), and regions (LOC).

IV. Experimental Setup (P)

A. Pre-training Setup

Using the continual (cont) method previously discussed, we put ConfliBERT-Spanish into practice. Similar to mBERT-Base, our architecture has 12 layers, 768 hidden units, 12 attention heads, and 110M parameters in total. We also use the same vocabulary files as mBERT and BETO. To train the models, we used two Nvidia A-100 GPUs, each with 10 GB of memory. We also employed an Adam optimizer [34] with the learning rate set to a peak value of 5e-5 and then linearly decaying. We trained the model using 512-byte sequences in order to accommodate the lengthy paragraphs of the new data. It took about three days to train each cont model.

B. Fine-Tuning Setup

ConfliBERT-Spanish fine-tunes the model using a task-specific layer on top of the pre-trained model and trains the entire architecture on a labelled dataset for downstream tasks. For the classification tasks, we needed a sequence classification/regression head on top of the combined BERT output. We used cross-entropy loss for BC and MLC. For the MLC tasks, we employed mean-square loss and fixed the discrimination thresholds at 0.5. For NER tasks, we predicted the sequence of BIO tags for each token in the input sentence; BIO tags are a popular format for tagging tokens in a chunking task. We also pre-processed the data to verify that it was in the proper CoNLL format [35]. For all datasets, we used a (60,20,20) split for training, testing, and development purposes.

BERT models use different casing to accommodate different tasks. Earlier work [20] utilized uncased models for some tasks and cased models for NER. Moreover, according to other studies [15], [17], uncased models for English performed slightly better than cased models in several domains, specifically when it comes to NER tasks. We decided to test mBERT and BETO and our ConfliBERT-Spanish variations in both cased and uncased variants to assess their performance in Spanish.

Using a single Nvidia A-100 GPU, we optimized our models over five epochs at a learning rate of 5e-05, a batch size of 16, and a maximum sequence length of 128 for NER, and 512 for both classification tasks. We ran all experiments ten times with different seeds. As performance measures, we employ F1 score for BC and for the MLC and NER tasks we use F1 macro, which is the average of F1 score of each class.

V. Results and Analysis

As Table II uses bold font to identify the best-performing model for each task. As the resutls show, ConfliBERT-Spanish consistently exhibited superior performance compared to both mBERT and BETO across all tasks. ConfliBERT-Spanish consistently achieved higher F1 scores for BC and macro F1 scores for MLC and NER in all cases, establishing its effectiveness and outperforming the baseline models.

1) BC Results: For all cases in BC, ConfliBERT-Spanish showed improved performance compared to the mBERT and BETO baseline models. In the BC evaluation on the Huffingtonpost Spanish dataset, ConfliBERT-Spanish based on mBERT-cased showed the best performance by reaching an F1 score of 89.60. In the BC task for the protest dataset, ConfliBERT-Spanish based on BETO-uncased performed best by reaching an F1 score of 87.25.

2) MLC Results: Our ConfliBERT-Spanish model consistently demonstrates performance improvements across all scenarios in MLC compared to the baseline models. In the evaluation of MLC using the InsightCrime dataset, the ConfliBERT-Spanish model based on mBERT-cased reported the highest performance with a macro F1 of 77.74. In the task MLC conducted on the Protest dataset, the ConfliBERT-Spanish based on mBERT-uncased version outperformed the other models by achieving a macro F1 of 63.48.

3) NER Results: Finally, results show that ConfliBERT-Spanish consistently outperforms the baseline models across all cases in NER tasks. Our ConfliBERT-Spanish model showed higher F1 macro than those of baselines. Especially, the ConfliBERT-Spanish version based on BETO-uncased performed best in NER task by reaching a macro F1 of 83.96. Given that the MX-news dataset contains many conflict-related entities in political texts, it is possible to state that

TABLE II: Summary of F1 score results for fine-tuned model evaluation

Dataset	Domain	Task	Models	mBERT		BETO	
				cased	uncased	cased	uncased
Huffing topnpost	Politics	BC	Baseline	87.57	86.29	88.16	87.50
			ConfliBERT Spanish	**89.60**	88.90	88.97	88.54
Protest	Conflict	BC	Baseline	79.56	83.64	82.95	85.54
			ConfliBERT Spanish	84.01	83.91	82.96	**87.25**
Insight Crime	Crime	MLC	Baseline	74.49	72.35	75.78	75.48
			ConfliBERT Spanish	**77.74**	77.13	77.31	76.15
Protest	Conflict	MLC	Baseline	56.49	46.88	58.07	58.10
			ConfliBERT Spanish	57.99	**63.48**	59.73	59.64
Mx News	Politics	NER	Baseline	82.92	82.69	83.36	78.72
			ConfliBERT Spanish	83.27	83.31	83.60	**83.96**

our ConfliBERT-Spanish model works well in detecting and classifying complex NER tasks in the conflict domain.

VI. Conclusion and Future Work

This research introduces ConfliBERT-Spanish, a domain-specific pre-trained language model for political conflict and violence in Spanish. Based on our previous work creating ConfliBERT in English [11], we advanced the multi-lingual capabilities of our cyberinfrastructure to study political violence by developing ConfliBERT-Spanish. For the pre-training step, we gathered and curated a significant domain-specific corpus from a variety of sources in Spanish. In the pre-training stage, we developed ConfliBERT-Spanish by implementing a continual training approach based on mBERT and BETO, the two most prevalent generic language models in Spanish.

To assess ConfliBERT-Spanish, we evaluate the performance of our pre-trained language model in various domain-specific NLP tasks using different datasets in Spanish specialized in the field of political conflict and violence. The results consistently show that our domain-specific ConfliBERT-Spanish outperforms generic mBERT and BETO in a variety of tasks including binary classification, multi-label classification, and named entity recognition in the relevant domain. These results are in line with other studies showing the value of generating domain-specific pre-trained language models that require specialized language for sophisticated tasks.

These results indicate the tremendous value of ConfliBERT-Spanish as a useful cyberinfrastructure tool to enhance the capacity of researchers and decision-makers in the conflict and security sectors who are interested in monitoring, evaluating, and forecasting political violence and conflict in Latin America. As several researchers, government agencies, and international organizations have noticed, Latin America is the most violent region in the world [1] as several countries are ravaged by political violence, organized crime, and social unrest. Having a language-specific and domain-specific cyberinfrastructure capable of analyzing vast volumes of text in Spanish with a high degree of accuracy will greatly increase the capacity of scholars to understand the dynamics of political violence and support government authorities in designing effective policies to address security challenges.

Our future research will focus on five areas. Given the scarcity of domain-specific text in Spanish, this development used a relatively small corpus to pre-train ConfliBERT-Spanish. As a first task, we will increase the size of our domain-specific corpora by combining web scraping, automated translation, and text augmentation. Second, we will expand hyper-parameters such as vocabulary size and epochs to improve ConfliBERT-Spanish performance, and we will pre-train ConfliBERT-Spanish from scratch. Third, we will evaluate ConfliBERT-Spanish on more difficult tasks such as comprehension, inference, question-answering, and uncertainty qualifying. To do so, we will develop our own training and evaluation databases, thus increasing the availability of domain-specific resources in Spanish. Fourth, we will develop specific applications for analyzing armed actors and organized criminal groups in Latin America. Finally, thanks to NSF support [OAC-2311142], the research team will help develop a community of ConfliBERT-Spanish users in Latin America. This community-building effort will enable scholars and government authorities to use the ConfliBERT-Spanish cyberinfrastructure to address pressing security challenges in Latin America.

Acknowledgments

This research was supported in part by NSF awards DGE-2039542, OAC-1828467, OAC-1931541, OAC-2311142, and DGE-1906630, ONR-N00014-20-1-2738, Army Research Office W911NF2110032, and IBM faculty award (Research). We thank the High Performance Computing (HPC) resources supported by the University of Arizona TRIF, UITS, and Research, Innovation, and Impact (RII), and maintained by the UArizona Research Technologies department. This work used Delta GPU at the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign through allocation CIS220162 from the Advanced

Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by NSF grants 2138259, 2138286, 2138307, 2137603, and 2138296.

REFERENCES

[1] United Nations Office on Drugs and Crime, "Global Study on Homicide 2019," Tech. Rep., 2019.

[2] 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.

[3] J. Osorio, V. Pavon, S. Salam, J. Holmes, P. T. Brandt, and L. Khan, "Translating CAMEO verbs for automated coding of event data," *International Interactions*, vol. 45, no. 6, pp. 1049–1064, 2019.

[4] J. Osorio, M. Mohamed, V. Pavon, and B.-O. Susan, "Mapping Violent Presence of Armed Actors," *Advances in Cartography in GIScience of the International Cartographic Association*, pp. 1–16, 2019. [Online]. Available: https://www.adv-cartogr-giscience-int-cartogr-assoc.net/1/16/2019/

[5] T. Anders, "Territorial control in civil wars: Theory and measurement using machine learning," *Journal of Peace Research*, vol. 57, no. 6, pp. 701–714, Nov. 2020, publisher: SAGE Publications Ltd. [Online]. Available: https://doi.org/10.1177/0022343320959687

[6] S. Bazzi, R. A. Blair, C. Blattman, O. Dube, M. Gudgeon, and R. Peck, "The promise and pitfalls of conflict prediction: evidence from colombia and indonesia," *Review of Economics and Statistics*, vol. 104, no. 4, pp. 764–779, 2022.

[7] J. Osorio and A. Beltran, "Enhancing the detection of criminal organizations in mexico using ml and nlp," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–7.

[8] F. Arias, A. Guerra-Adames, M. Zambrano, E. Quintero-Guerra, and N. Tejedor-Flores, "Analyzing Spanish-Language Public Sentiment in the Context of a Pandemic and Social Unrest: The Panama Case," *International Journal of Environmental Research and Public Health*, vol. 19, no. 16, p. 10328, Aug. 2022.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[10] L. García Montenero, "El Español: Una Lengua Viva. Informe 2022," Instituto Cervantes, Tech. Rep., 2022. [Online]. Available: https://cervantes.org/es/sobre-nosotros/publicaciones/el-espanol-una-lengua-viva-informe-2022/

[11] Y. Hu, M. Hosseini, E. S. Parolin, J. Osorio, L. Khan, P. Brandt, and V. D'Orazio, "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*, 2022, pp. 5469–5482.

[12] S. Alsarra, L. Abdeljaber, W. Yang, N. Zawad, L. Khan, P. Brandt, J. Osorio, and V. D'Orazio, "Conflibert-arabic: A pre-trained arabic language model for politics, conflicts and violence," in *Proceedings of the 2023 International Conference on Recent Advances in Natural Language Processing (RANLP 2023)*, 2023.

[13] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez, "Spanish pre-trained bert model and evaluation data," in *PML4DC at ICLR 2020*, 2020.

[14] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, 09 2019. [Online]. Available: https://doi.org/10.1093/bioinformatics/btz682

[15] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.

[16] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The muppets straight out of law school," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Nov. 2020, pp. 2898–2904.

[17] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3615–3620.

[18] L. B. Barbecho, "Bert for humanists," 2023.

[19] Y. Jin, E. Jang, J. Cui, J.-W. Chung, Y. Lee, and S. Shin, "DarkBERT: A Language Model for the Dark Side of the Internet," *arXiv*, no. 2305.08596, 2023. [Online]. Available: https://arxiv.org/abs/2305.08596

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.

[21] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," *arXiv preprint arXiv:2003.00104*, 2020.

[22] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4996–5001. [Online]. Available: https://aclanthology.org/P19-1493

[23] P. A. Schrodt, "Cameo: Conflict and mediation event observations event and actor codebook," *Pennsylvania State University*, vol. 610, p. 35, 2012.

[24] A. Eisele and Y. Chen, "Multiun: A multilingual corpus from united nation documents." in *LREC*, 2010.

[25] J. Tiedemann, "Parallel data, tools and interfaces in opus." in *Lrec*, vol. 2012. Citeseer, 2012, pp. 2214–2218.

[26] E. S. Parolin, M. Hosseini, Y. Hu, L. Khan, J. Osorio, P. T. Brandt, and V. D'Orazio, "Multi-CoPED: A Multilingual Multi-Task Approach for Coding Political Event Data on Conflict and Mediation Domain," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022.

[27] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018, pp. 353–355.

[28] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Superglue: A stickier benchmark for general-purpose language understanding systems," *Advances in neural information processing systems*, vol. 32, 2019.

[29] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for squad," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 784–789.

[30] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, "Race: Large-scale reading comprehension dataset from examinations," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 785–794.

[31] Mendonça, D. Jaquette, D. Graff, and D. DiPersio, "Spanish Gigaword Third Edition LDC2011T12. Web Download." 2011. [Online]. Available: https://catalog.ldc.upenn.edu/ldc2011t12

[32] E. Skorupa Parolin, M. Hosseini, Y. Hu, L. Khan, P. T. Brandt, J. Osorio, and V. D'Orazio, "Multi-coped: A multilingual multi-task approach for coding political event data on conflict and mediation domain," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022, pp. 700–711.

[33] O. Ramos-Flores, D. Pinto, M. Montes-y Gómez, A. Vázquez, D. Pinto, V. Singh, and F. Perez, "Probabilistic vs deep learning based approaches for narrow domain ner in spanish," *J. Intell. Fuzzy Syst.*, vol. 39, no. 2, p. 2015–2025, jan 2020. [Online]. Available: https://doi.org/10.3233/JIFS-179868

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[35] E. F. Sang and F. De Meulder, "Introduction to the conll-2003 shared task: Language-independent named entity recognition," *arXiv preprint cs/0306050*, 2003.