PRE-TRAINED SPANISH LANGUAGE MODEL FOR POLITICAL CONFLICT AND VIOLENCE

by

Wooseong Yang

APPROVED BY SUPERVISORY COMMITTEE:

___

Latifur Khan, Chair

___

Lawrence Chung

___

Xinya Du

This thesis is dedicated to my family

for their endless support and love.

PRE-TRAINED SPANISH LANGUAGE MODEL FOR
POLITICAL CONFLICT AND VIOLENCE

by

WOOSEONG YANG, BS, MS

THESIS

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN

COMPUTER SCIENCE

THE UNIVERSITY OF TEXAS AT DALLAS

August 2023

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Dr. Latifur Khan, for his invaluable guidance throughout my thesis journey. His constructive feedback and encouragement have been instrumental in shaping my research.

Additionally, I am incredibly grateful to Dr. Javier Osorio for his close guidance and support. His mentorship and expertise played a significant role, and I am truly fortunate to have had the opportunity to work under his supervision.

I would like to extend my sincere thanks to my parents and two older sisters for their unwavering love and encouragement. Their belief in me and constant encouragement have been a driving force behind my accomplishments.

Lastly, I would like to express my gratitude to all the individuals around me who have provided their support, whether it be friends, colleagues, or mentors. Their words of encouragement, motivation, and assistance have been invaluable, and I am deeply thankful for their presence in my life.

May 2023

PRE-TRAINED SPANISH LANGUAGE MODEL FOR

POLITICAL CONFLICT AND VIOLENCE

Wooseong Yang, MSCS
The University of Texas at Dallas, 2023

Supervising Professor: Latifur Khan, Chair

Examining political conflict and violence remains a persistent challenge for the political science and policy communities, because there comes large amount of text to be dealt with to monitor political conflict and violence. In order to contribute to the advance of conflict research in Spanish speaking society, we introduce ConfliBERT Spanish, a domain-specific pre-trained language model tailored for Spanish political conflict and violence analysis. Our method begins with the collection of a comprehensive domain-specific corpus from diverse sources, which is then utilized for language modeling purposes. ConfliBERT Spanish is subsequently developed using continual pre-training process. To evaluate the practical performance of ConfliBERT Spanish, we assembled 5 datasets and implemented 3 tasks using them. Through multiple experiments and evaluations on various versions of ConfliBERT Spanish, we proved that ConfliBERT Spanish outperforms in analyzing Spanish political conflict and violence compared to BERT baseline models.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The political conflict and violence has been one of the main concern of political scientists in academia and policy communities (Jacoby, 2007). The conflict research is a sub-field of political science study that examines the causes, dynamics, and consequences of conflict. It covers a wide range of topics, including violence, protest, repression, terrorism, human rights abuses, genocide, and war. The goal of conflict research is to develop a better understanding of the domain so that it can be prevented or managed more effectively. The research can be utilized to policy making and educating the causes and consequences of conflict. For convenience, we will abbreviate political conflict and violence as ”conflict” from now.

The emergence and advance of data mining techniques (Yen et al., 2002; Al-Naami et al., 2016; Khan and McLeod, 2000; Parveen et al., 2011; Thuraisingham et al., 2008; Abrol and Khan, 2010a; Awad et al., 2008) have significantly impacted various fields (Luo et al., 2007; Sahs and Khan, 2012; Hamlen et al., 2010; Ayoade et al., 2018; Luo et al., 2004; Parveen et al., 2011; Abedin et al., 2006; Shaon et al., 2017; Masud et al., 2007; Abrol and Khan, 2010b; Wang et al., 2004; Tu et al., 2008; Wang and Khan, 2006; Awad and Khan, 2007), including political conflict and violence domain. Data mining techniques involve extracting useful patterns and insights from large datasets, enabling researchers to uncover hidden relationships and gain a deeper understanding of complex phenomena (Khan et al., 2007; Masud et al., 2010; Abrol et al., 2015; Osman, 2019; Masud et al., 2008; Haque et al., 2016; Golnabi et al., 2006; Haque et al., 2016; Masud et al., 2011; Awad et al., 2004; Breen et al., 2002; Petrushin and Khan, 2007; Wang and Khan, 2006; Nessa et al., 2008). In conflict research, these techniques offer potential for analyzing vast amounts of data related to political conflicts and violence. By applying data mining techniques to conflict research, researchers can identify conflict patterns, root causes, escalation factors, and consequences. They can explore the interplay of various factors, such as social, economic, and political variables, to develop more comprehensive models and theories of conflict. Additionally, data mining can aid in predicting and forecasting conflict events, identifying early warning signs, and assessing the effectiveness of conflict management strategies. Ultimately, the application of data mining techniques in conflict research holds great promise in enhancing our knowledge and providing valuable insights for policymakers, enabling more informed decision-making processes and contributing to the prevention and resolution of conflicts.

In the early days, conflict researchers manually coded to track conflict events around the world (Raleigh et al., 2010). However, the method is time-consuming and it is not always possible to keep up with the pace of rapidly changing conflicts. Also, manual coding often focuses on specific types of conflict events between particular types of actor entities (Sundberg and Melander, 2013). In recent years, there has been a shift towards using data mining based automated methods for tracking conflict events (Bond et al., 2003; O'brien, 2010; Osorio and Reyes, 2017; Schrodt and Hall, 2006; Alliance, 2015; Norris et al., 2017; Lu and Roy, 2017; Ward et al., 2013). The automated systems have capacity to encompass various conflict and cooperation events that involve numerous political actors. Furthermore, these systems are capable of extracting large amounts of data that exceed the capacity of manual coding efforts. For example, the Integrated Crisis Early Warning Systems utilizes automated event data to forecast potential conflicts and conduct other types of political science research (Bagozzi et al., 2021; Beger et al., 2016; Brandt et al., 2022).

However, automated methods also have limitations that they sometimes are not guaranteed to be accurate, and they may not be able to capture the context of human conflict. Existing automated systems for conflict research rely on pattern matching techniques and large dictionaries, which often yield low-accuracy results and are too costly to maintain.

Although recent efforts by political scientists have employed traditional machine learning and deep learning techniques to analyze political conflict and violence (Hanna, 2017; Osorio et al., 2020; Beieler, 2016; Glavaš et al., 2017; Parolin et al., 2020), standard supervised learning requires labeled data, which is expensive to obtain due to the expertise required for quality annotation. This led political scientists to use Natural Language Processing (NLP) techniques in the conflict field. By using NLP techniques, conflict scholars can develop more accurate and efficient automated systems for tracking conflict.

NLP is a field of computer science that deals with understanding and interpreting human natural languages (Chowdhary and Chowdhary, 2020). NLP techniques can be used to extract structured information from text, such as the entities, events and relationships that are mentioned in an article.

Among many NLP techniques, especially pre-trained language models have been shown to be effective on most of the NLP tasks (Vaswani et al., 2017a; Devlin et al., 2018; Liu et al., 2019; Yang et al., 2019; Radford et al., 2019a; Brown et al., 2020a; Meta, 2023). The success of pre-trained language models have three major reasons. First, the availability of large-scale unlabeled text data has made it possible to train these models on a massive scale. Second, the development of powerful computational devices has made it possible to train these models in a reasonable amount of time. Third, the introduction of extensive benchmarks has allowed researchers to compare the performance of different models on a various tasks. As a result of these advances, the models are now widely applied in NLP research.

Many language models are pre-trained on general-domain corpora, such as Wikipedia, BookCorpus, and WebText (Zhu et al., 2015; Radford et al., 2019a). However, recent research has shown that pre-training on domain-specific corpora can improve the performance of the model on those domains (Lee et al., 2020; Beltagy et al., 2019; Alsentzer et al., 2019; Lewis et al., 2020; Gu et al., 2021; Chalkidis et al., 2020; Hu et al., 2022). These models have been shown to be effective at tasks such as natural language inference and question answering.

In the field of conflict research, a model called ConfliBERT (Hu et al., 2022) has been proposed and shown to be effective. However, ConfliBERT has a limitation that it only can be applied on English text because it was trained on English corpus. We decided to extend ConfliBERT to multilingual setting, starting with Spanish. Spanish is one of the most spoken languages in the world. Also, the political situation in Spanish-speaking countries, especially in Latin America, is getting more serious. Thus, political scientists in Spanish-speaking countries need tools to help them analyze conflict related text. This is the main reason why we started with Spanish, so the Spanish ConfliBERT has potential to be used by Spanish-speaking conflict researchers to analyze and manage political conflicts and violence.

We propose ConfliBERT Spanish that is a pre-trained language model specifically tailored for research on Spanish conflict and political violence. It was developed by collaboration of conflict scholars and computer scientists, and it is designed to improve performance on conflict research tasks while also reducing the need for manual work.

Our work provides the following key contributions.

- We curate a substantial corpus specifically tailored for Spanish language modeling within the domains of political violence, conflict, cooperation, and diplomacy.

- Leveraging our domain-specific corpora, we develop ConfliBERT Spanish, a pre-trained language model that is made publicly accessible, directly benefiting the political scientists and policy communities.

- To assess the practical applicability of our model, we compile 4 datasets and conduct 5 tasks that are highly relevant to conflict research. This comprehensive evaluation of language models for Spanish conflict studies is the first of its kind.

- We thoroughly evaluate different versions of ConfliBERT Spanish and demonstrate its superior performance compared to models trained on generic domains. Furthermore, we conduct analysis of various tasks to understand the results.

CHAPTER 2

BACKGROUND

2.1 Language Representation Learning

Language representation learning is an unsupervised learning method that aims to learn a general-purpose representation of language that can be used across various NLP tasks. The primary goal is to extract useful features, implicit linguistic principles and common sense knowledge from textual data, such as lexical meanings, syntactic structures, semantic roles, and pragmatics (Qiu et al., 2020; Latif et al., 2020).

The initial stage of language representation is to convert discrete language symbols into a distributed embedding space, which is called word embedding (Mikolov et al., 2013a). Word embedding can be classified into two categories: non-contextual embedding and contextual embedding.

2.1.1 Non-contextual Embedding

Non-contextual embedding is static embedding of a word that does not change depending on the context in which it appears. Non-contextual word embedding, such as word2vec (Mikolov et al., 2013b) and Glove (Pennington et al., 2014a), are typically learned by functions that map each word type to a single vector. The resulting embedding represent each word as a dense vector of fixed length, and can be used as input feature for a wide range of NLP tasks, such as sentiment analysis (Hussein, 2018) or text classification (Kowsari et al., 2019).

Non-contextual embedding can be computationally efficient and can be used with small amounts of training data. However, non-contextual embedding has limitation in capturing the meaning of words in different contexts. Since it is fixed, it cannot capture the variability of word meaning which is different depending on the context in which they appear. To overcome this limitation, contextual embedding has been developed.

2.1.2 Contextual Embedding

Contextual embedding is dynamic representation of words that vary depending on the context in which they are used. Contextual embedding, such as ELMO (Peters et al., 2018), BERT (Devlin et al., 2018), XLNET (Yang et al., 2019), GPT-3 (Brown et al., 2020a) are typically learned by a functions that map each word type to different vectors depending on the context. The resulting embedding captures not only the meaning of a word, but also its relationship to the other words within the document, making them particularly useful for tasks that require understanding of document-level context.

Contextual embedding has gained popularity in recent years, particularly with the advent of large pre-trained language models such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) and GPT-3 (Brown et al., 2020a). These models have demonstrated state-of-the-art performance on a wide range of NLP tasks, and have led to significant advances in the field of NLP.

2.2 Pre-trained Language Models

With the development of deep learning, the size of language model parameters that is used for contextual embedding has increased rapidly. Accordingly, much larger dataset is needed to fully train model parameters and avoid overfitting. However, building large-scale labeled dataset is a challenging task as it costs too much annotation cost.

Unsupervised methods have been proposed to tackle the issue as it is relatively easy to build large-scale unlabeled corpora. To manage the large unlabeled text data effectively, the strategy is to learn representation first and then use the representation for tasks. This process is called pre-training and the Pre-trained Language Models (PLMs) has been proved by many researches to be effective method to get contextual embedding (Qiu et al., 2020).

PLMs have been improved rapidly in last decade. In the early years, Word2Vec (Mikolov et al., 2013b) that trains the fixed sized word vector has been proposed. Word2Vec uses two types of algorithms to generate word vectors which are Skip-Gram and Continuous Bag of Words (CBoW) (Mikolov et al., 2013). Word2Vec has been widely used in various NLP tasks, such as text classification and sentiment analysis. However, Word2Vec has a limitation that it cannot completely cover the data in the corpus as its dimension is small. Global Vectors for Word Representation (GloVe) (Pennington et al., 2014b) was proposed to overcome the limitation, using co-occurrence counts to capture the global patterns of words in corpus. Both Word2Vec and GloVe are non-contextual as they use fixed representation for same word. Therefore, these models are not able to represent complex context.

Embedding from Language Models (ELMo) (Peters et al., 2018) was proposed to overcome the raised problem. ELMo generates contextualized embeddings that capture the meaning of a word in the context. It extracts embeddings from a bi-directional LSTM pre-trained on corpus. This enables ELMo to learn context of the corpus, having different embeddings depending on the context in which it appears. Even though ELMo is based on bi-directional LSTM, it is still one-way language model. This characteristic limit its ability to model semantic information of corpus. To address it, Google AI proposed pre-trained Bidirectional Encoder Representations from Transformers (BERT).

2.3 BERT

To overcome the limitations of ELMo, BERT (Devlin et al., 2018) uses a bidirectional approach in pre-training phase, enabling the model to learn contextual relationships between words in a sentence by considering both left and right context. BERT uses two strategies for deep two-way joint training which are Masked Language Model (MLM) and Next Sentence Prediction (NSP). In the MLM strategy, some words in a sentence are randomly masked and the model is trained to predict the original word according to the remaining words. This strategy helps the model to learn the relationships between different words in a sentence and to develop a deeper understanding of language. In the NSP strategy, the model is given pairs of sentences and is trained to predict whether the second sentence is a continuation of the first sentence of not. The strategy helps the model to capture the relationship between models and develop a sense of coherence and continuity in natural language.

The emergence of BERT has had a significant impact and promoted the development of the NLP field. It has led to the development of other BERT-based PLMs. RoBERTa (Robustly Optimized BERT approach) (Liu et al., 2019) is a variant of BERT that uses a larger dataset, that changed the masking method from static to dynamic and canceled to use NSP to achieve better performance. ALBERT (A Lite BERT) (Lan et al., 2019) uses a technique that shares parameters cross-layer to reduce the number of parameters while maintaining performance. DistilBERT (Distilled version of BERT) (Sanh et al., 2019) is designed to be more efficient and faster to train and deploy than the original model. ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) (Clark et al., 2020) utilizes RTD (Replaced Token Detection) instead of MLM to solve the inconsistency between the pre-training and the fine-tuning of mask.

2.4 Domain-specific BERT

Although many language models are built using general domain corpora, recent research indicates that pre-training on domain-specific corpora can enhance the performance of downstream tasks on the domain (Lee et al., 2020; Beltagy et al., 2019; Chalkidis et al., 2020; Hu et al., 2022). BioBERT (Lee et al., 2020) is a BERT-based model that is pre-trained on bio-medical literature and fine-tuned on downstream tasks, such as bio-medical named entity recognition and bio-information extraction, and showed good results. SciBERT (Beltagy et al., 2019) is a BERT-based model that is pre-trained on scientific publications and fine-tuned on downstream tasks, such as scientific article classification and scientific question answering, showing better results compared to BERT. LegalBERT (Chalkidis et al., 2020) is a BERT-based model that is pre-trained on legal documents and fine-tuned on downstream tasks, such as legal document classification and legal named entity recognition, showing better results than original BERT. ConfliBERT (Hu et al., 2022) is pre-trained on a large corpus of texts that contain conflicting information, such as news articles, social media posts and Wikipedia. Then, it is fine-tuned on downstream tasks, such as conflict text classification and political named entity recognition, showing improved results.

CHAPTER 3

CONFLIBERT SPANISH

As stated in previous chapter, BERT demonstrated its competitive performance among pre-trained language models across different natural language processing tasks. To utilize the advantage in the conflict domain, a preliminary study proposed ConfliBERT which is pre-trained on large corpus of English conflict text. In this study, we apply this approach to Spanish to support Spanish-speaking researchers around world by enabling the analysis of political conflicts. First part of this chapter briefly explains basic methods for PLMs: Long-Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), Attention (Bahdanau et al., 2015), and Transformer (Vaswani et al., 2017b). Latter part explains key concepts of BERT: bi-directional structure, pre-training and fine-tuning.

3.1 Basic Methods for PLMs

LSTM Recurrent Neural Networks (RNNs) (Rumelhart et al., 1986) have been commonly used for processing sequential data. but they have an critical shortage called vanishing gradient problem. When dealing with long sequential data, it becomes challenging to preserve earlier information for latter stages, resulting gradient disappearance. To address this issue, LSTM (Hochreiter and Schmidhuber, 1997) has been proposed. LSTM is an enhanced RNN model that has input, forget, and output gates to process and preserve information. The forget gate controls how much of information from the previous unit to be preserved for the current unit, while the input gate decides how much of the immediate status can be input of the unit status. Lastly, the output gate determines how much of the unit status can be used as the present output value of the LSTM. The architectures of RNN and LSTM are depicted in Figure 3.1 and Figure 3.2.

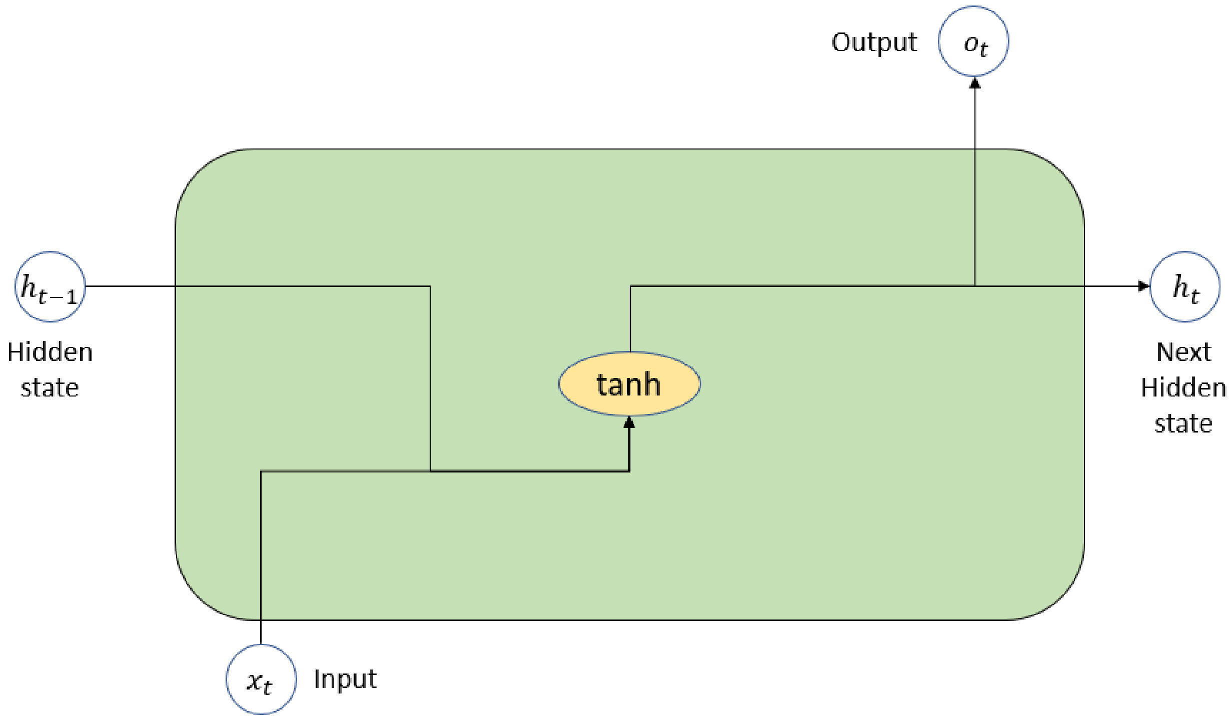

Figure 3.1. Architecture of Recurrent neural network (RNN) model

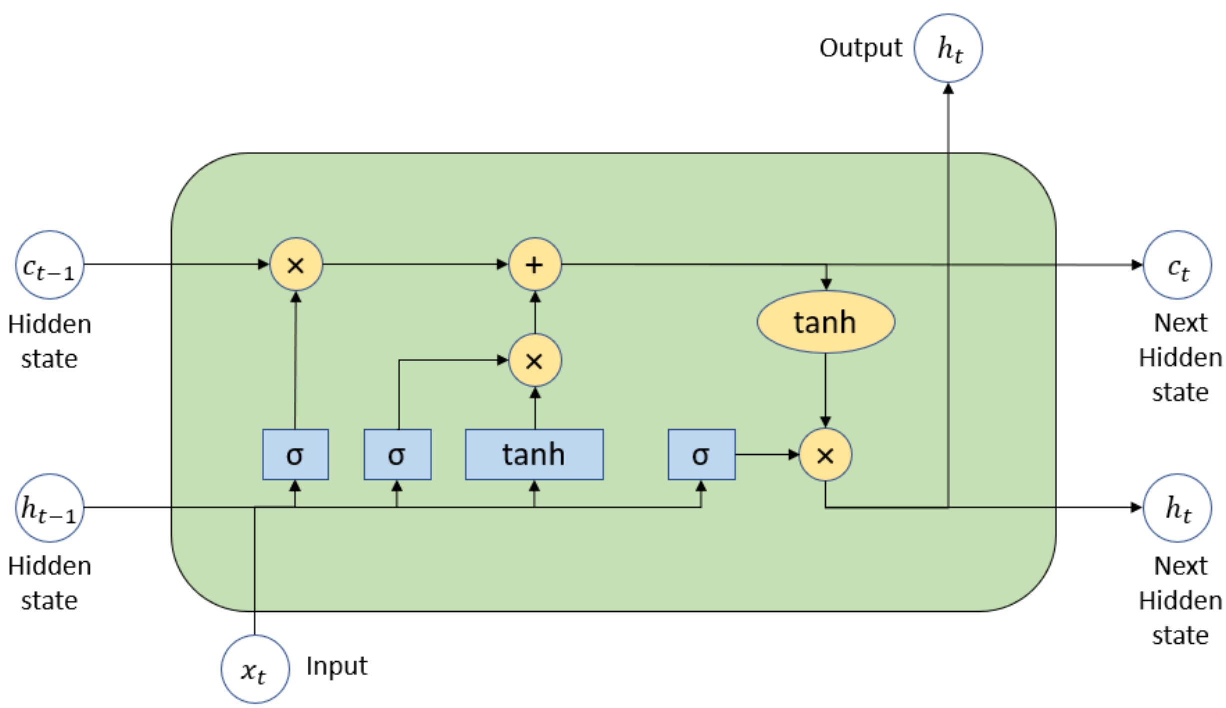

Figure 3.2. Architecture of Long short-term memory network (LSTM) model

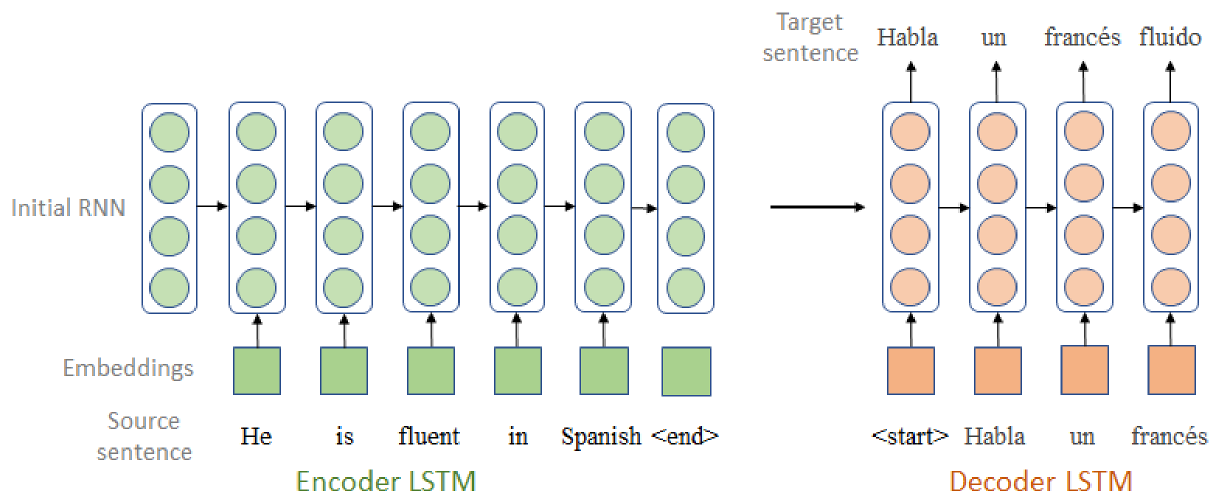

Figure 3.3. Basic architecture of seq2seq model

Encoder-Decoder Encoder-decoder is a standard modeling framework for sequence-to-sequence (seq2seq) tasks (Sutskever et al., 2014). This framework consists of two components: encoder and decoder. Encoder takes information of source sequence as input and makes its representation as output. Decoder uses the representation as input to generate target sequence. The representative application of encoder-decoder is seq2seq model which consists of two LSTMs, one for the encoder and another for the decoder. Encoder LSTM reads the source sentence and the final state becomes output representation of it. The decoder generates the target sentence based on this representation which is called context vector. As encoder compresses the source sentence into a single vector, the seq2seq model faces bottleneck problem that the model fails to preserve enough information and forgets.

Attention The attention mechanism (Bahdanau et al., 2015) was proposed to overcome the aforementioned shortage of traditional RNNs and LSTMs. It assigns weights to all hidden states in the encoder and feeds the weighted sum of these states to the decoder layer, enabling it to concentrate more on inputs that are relevant to the current task. Instead of

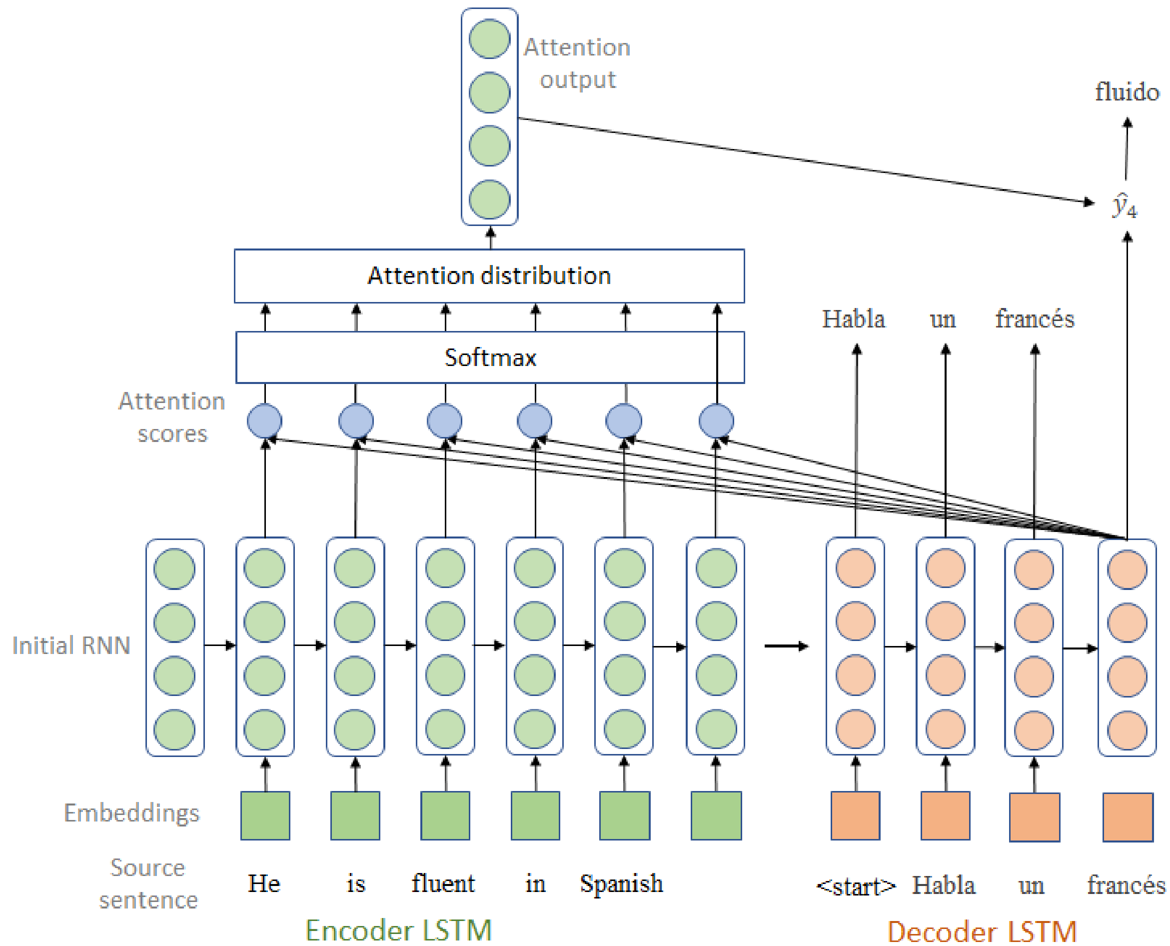

Figure 3.4. Architecture of seq2seq model with attention

processing input sequentially, the attention mechanism allows the model to focus on specific parts of the input.

Later, a self-attention was proposed where the model attends to its own input sequence instead of a different sequence. The model computes a weighted sum of input hidden states based on the relationships between each token and other tokens in the same sequence. Self-attention enables the model to capture long-range dependencies and the relationships between different parts of the input sequence. Furthermore, self-attention mechanism replaced the loop layer in the encoder-decoder architecture with multi-headed self-attention, resulting in a significant improvement in training speed. While the seq2seq model with attention and self-attention have been proposed, the models retains the drawbacks that come from the use of LSTM.

Transformer The transformer uses self-attention extensively to circumvent the drawbacks (Vaswani et al., 2017a). The transformer circumvent the drawbacks by employing a self-attention mechanism to encode and process input sequences, without requiring recurrent connections. The transformer architecture consists of an encoder and a decoder, each of which consists of multiple layers of self-attention and feedforward neural networks. In the encoder, the self-attention mechanism is used to capture the relationships between different parts of the input sequence, while in the decoder, it is used to attend to relevant parts of the encoder's output during the generation of the output sequence. Furthermore, the model can leverage the power of deep neural networks to improve efficiency and performance. The transformer has become the backbone of many state-of-the-art NLP models, especially PLMs, such as BERT family, GPT family, and T5 (Devlin et al., 2018; Lan et al., 2019; Liu et al., 2019; Radford et al., 2018, 2019b; Brown et al., 2020b; Raffel et al., 2020). The application of self-mechanism allows it to efficiently capture long-range dependencies in sequences, making it appropriate for many NLP tasks such as language modeling, text classification, and more.

3.2 Key Concepts of BERT

Among the transformer-based PLMs, BERT has shown good performance in various NLP tasks. There are several key concepts of BERT behind the success.

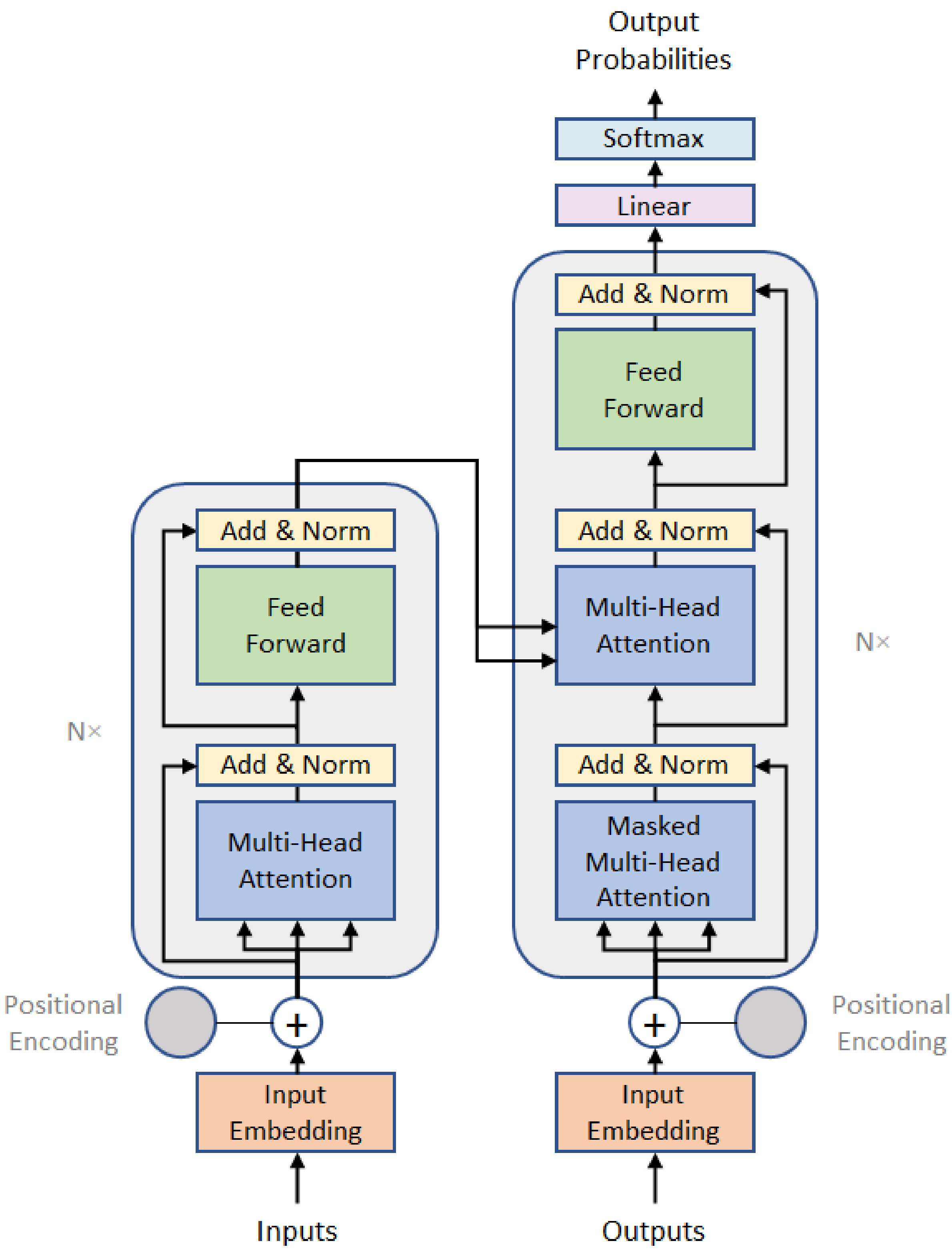

Figure 3.5. Architecture of Transformer model

Bi-directional Structure The bi-directional structure is a key element of the BERT architecture. Unlike other pre-trained language models that only deals text in one direction, BERT deals text in both directions, allowing it to capture deeper context of each word in a sentence. The two-way learning is achieved by a multi-layer bidirectional transformer encoder, which is a type of neural network architecture that uses self-attention to encode the input sequence.

Pre-training BERT is pre-trained on a large corpus of text using two unsupervised learning architectures: MLM and NSP. In MLM, a certain portion of words in the input texts are masked and the model is trained to predict the masked words based on the context of the nearby words. The strategy enables BERT to capture the meaning of words in their specific context. In NSP, the model is trained to predict whether two sentences in the input text are sequential or not. This way of training enables BERT to capture the semantic relationships between sentences. In our work, we use MLM to pre-train our own BERT model.

In the context of pre-training BERT on our own data, there are two strategies: learning from scratch (scr) and continual learning (cont). The scr strategy refers to pre-training a BERT model on text data, starting with randomly initialized parameters. The strategy trains model from scratch, without using the information learned in existing BERT model. On the other hand, the cont strategy involves using an existing BERT and continuously pre-train it on additional domain-specific data. The strategy makes it possible to build domain-specific BERT with smaller set of domain-specific data. Because scr strategy requires enormously large data for pre-training, we used cont strategy which can be more efficient and effective in achieving good performance. Frameworks of cont and scr strategies are described in Figure 3.6 and Figure 3.7.

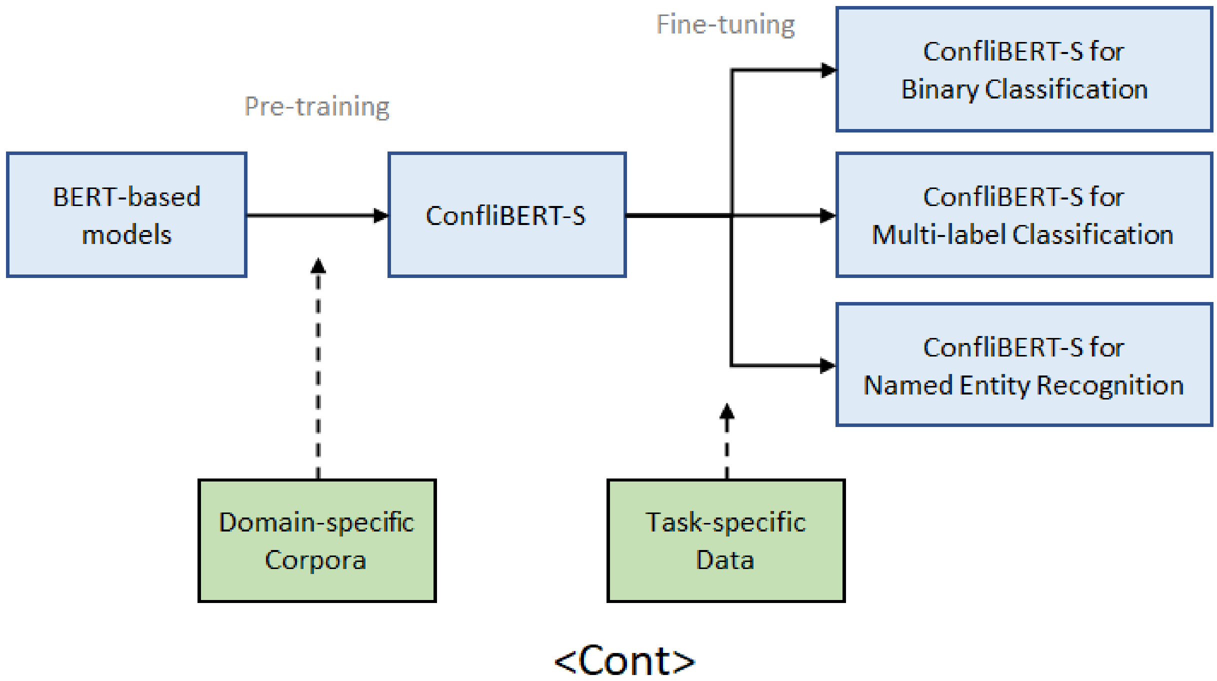

Figure 3.6. Framework of cont strategy

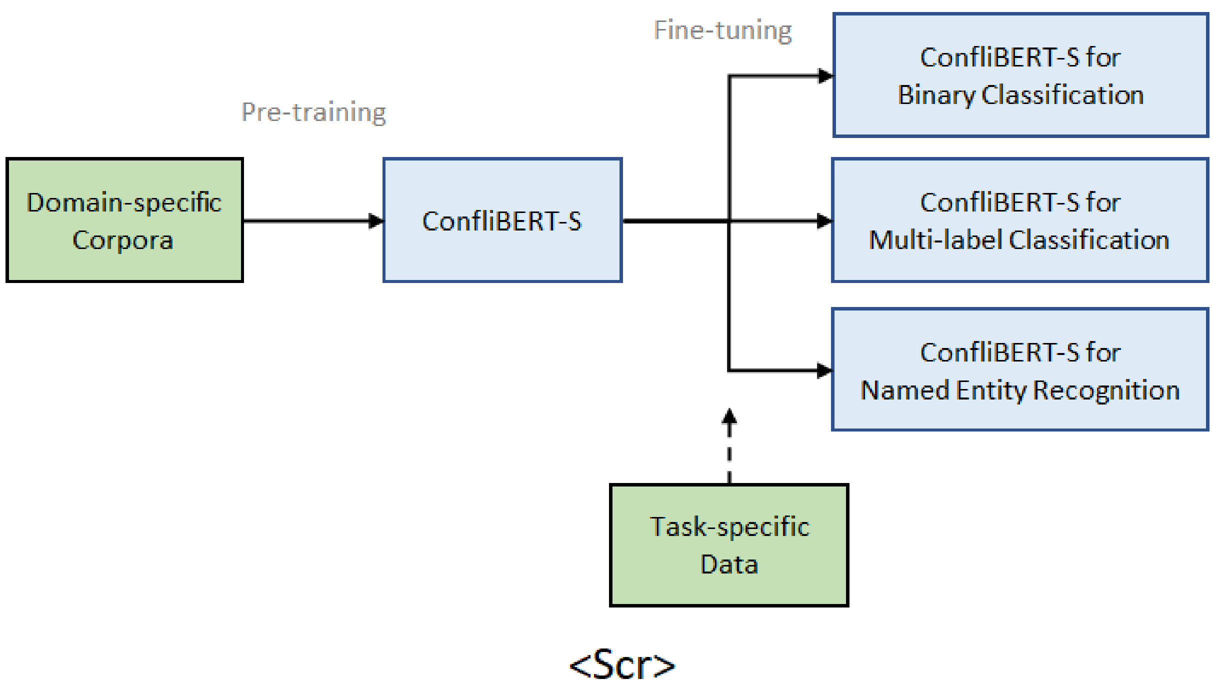

Figure 3.7. Framework of scr strategy

Fine-tuning After pre-training on domain-specific data, the BERT-based model can be fine-tuned on specific downstream tasks, such as text classification and named entity recognition.

In this work, we first pre-trained BERT on Spanish political conflict data following cont strategy to get ConfliBERT Spanish model. Then, we fine-tuned the model for downstream tasks (binary classification, multi-label classification, and named entity recognition) to demonstrate the advantage of the ConfliBERT Spanish. Details of the experiment process is explained on the following chapter.

3.2.1 ConfliBERT Spanish

As described above, BERT models using MLM strategy have achieved competitive performance among other transformer models in various NLP tasks. Furthermore, BERT has been applied to various domains and shown reasonable results. In this sense, the political conflict and violence domain requires domain-specific BERT in the aspect that the language model makes analyzing the related texts accurate and efficient. Although ConfliBERT (Hu et al., 2022) was proposed and successful, its usage is still limited to English. Therefore, we decided to develop a Spanish domain-specific BERT in political conflict and violent domain to expand the use of ConfliBERT to Spanish political science society.

Domain-specific pre-training We applied cont for adapting BERT to the conflict domain. The strategy starts training with existing checkpoint and vocabulary of BERT, and trains for additional steps on a domain-specific corpus. As BERT has already been pre-trained about a million steps on the general domain, cont requires fewer steps compared to scr (Lee et al., 2020). For this reason, we adopted the cont strategy in this work.

Figure 3.8. Example image of Spanish news website

Corpus for pre-training To develop ConfliBERT Spanish, the initial stage is to construct a domain-specific corpus for pre-training. To the best of my knowledge, there are only few public datasets which purely contains Spanish political conflict and violence texts. Therefore, we collected texts from news websites of Spanish speaking country and constructed political conflict and violence corpus. Details of the dataset are described in Experiment chapter. After constructing the corpus, we trained BERT-base model using Cont strategy to gain ConfliBERT Spanish model. The example of Spanish news website is shown in Figure 3.8 and that of crawled pre-training corpus is shown in Figure 3.9.

	news_outlet	title	date	text
0	abcspanish	Una milicia afín a Al Qaeda se responsabiliza ...	2009-07-30	Una milicia afín a Al Qaeda se responsabiliza...
1	abcspanish	Luna de miel con Mohamed VI	2009-07-30	De los tiempos de Aznar a los de Zapatero medi...
2	abcspanish	Los nicaragüenses temen la extensión del confl...	2009-07-30	La crisis política en Honduras y el protagoni...
3	abcspanish	Mohamed VI insiste en su plan de autonomía par...	2009-07-30	La televisión pública marroquí ha retransmi...
4	abcspanish	Las fuerzas nigerianas asaltan una mezquita y ...	2009-07-30	Las fuerzas de seguridad de Nigeria asaltaron ...
...	...	...	...	...

Figure 3.9. Example for crawled pre-training corpus

Fine-tuning for downstream tasks The next stage is fine-tuning the pre-trained model to each downstream tasks. During each fine-tuning process, the ConfliBERT Spanish is adapted to specific downstream tasks by adding a task-specific layer on top of the ConfliBERT Spanish and fine-tuning the weights of the entire model on a task-specific model. In this work, we applied ConfliBERT Spanish to three downstream tasks that seem to be useful for Spanish political scientists.

- Binary Classification: Binary classification is a type of classification problem where the objective is to classify the input into one or two classes. For example, in binary image classification, the object can be set to classify images as either dog or not dog. Another example can be sentiment analysis, where the objective is to classify the sentiment of a text as positive or negative.

- Multi-label Classification: Multi-label classification is a type of classification problem where the objective is to assign one or more labels to an input. For example, in multi-label image classification, the objective is to classify an image into multiple categories, such as "dog", "cat", "tree", and "bus".

- Named Entity Recognition (NER): NER is one type of NLP task where the objective is to identify and classify named entities in text. Named entities refer to objects, people, places, organizations and other items. For example, in the sentence "Barack Obama was president of the United States", the named entities can be "Barack Obama", "president", and "United States".

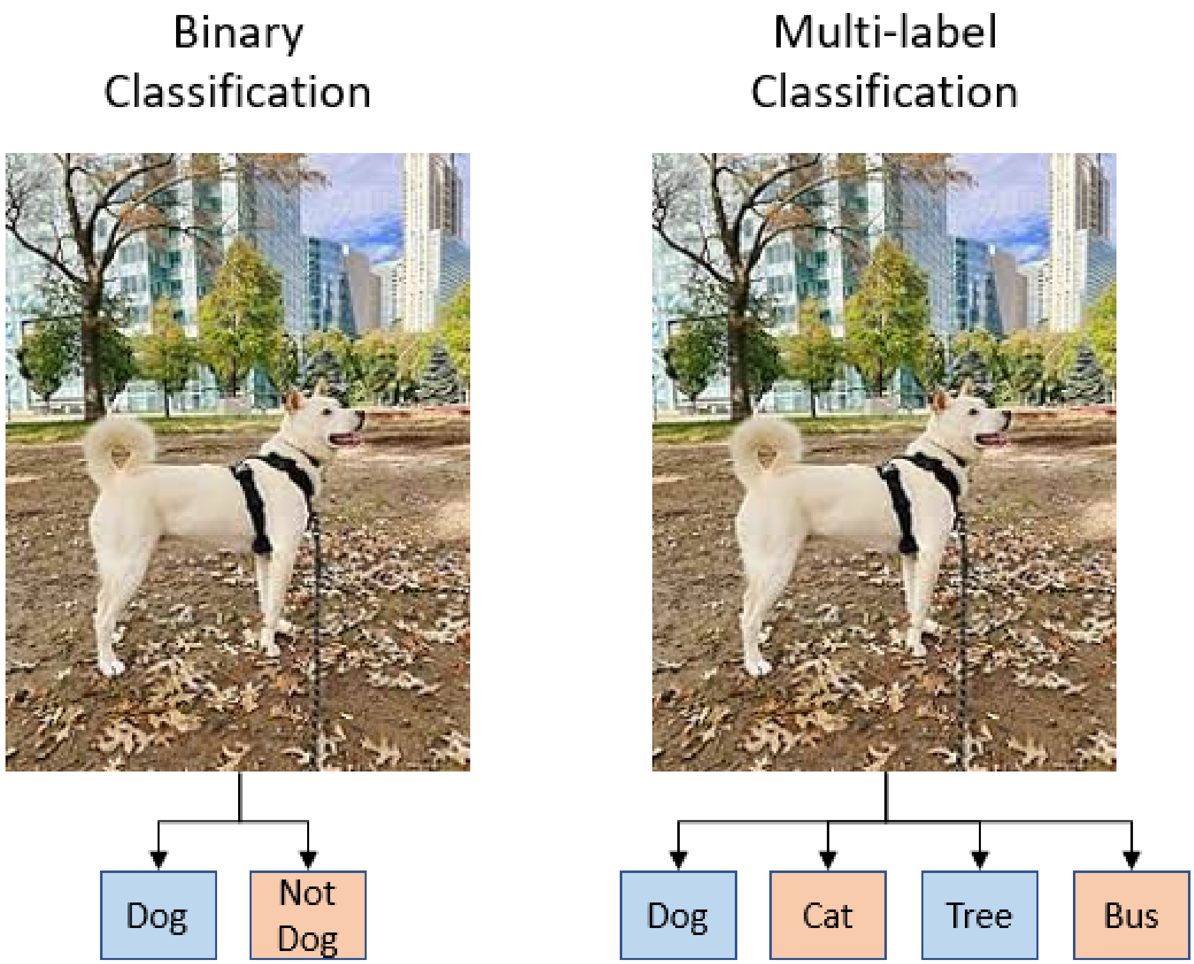

Figure 3.10. Example for comparison between classification tasks

CHAPTER 4

EXPERIMENT

A large amount of text data is required to train a BERT-based model as illustrated in Table 4.1. There are several Spanish corpus open to the public, such as Oscar (Abadji et al., 2022). However, due to the nature of our task that requires a corpus related to political conflict and violence, there is limitation in using the generic Spanish corpus. Therefore, we constructed our own Spanish corpus by crawling conflict-related texts from news websites.

4.1 Experiment Setup

Pre-training Setup We implemented ConfliBERT Spanish using the aforementioned continual (cont) techniques. The architecture used is based on Multilingual BERT (Devlin et al., 2018) which utilizes 12 layers, 768 hidden units, 12 attention heads, and a total of 110M parameters. We used the vocabulary file of the original Multilingual BERT (Devlin et al., 2018) and BETO (Cañete et al., 2020). We used 2 Nvidia A-100 GPUs with 10GB memory to train the models. We used an Adam optimizer (Kingma and Ba, 2015) with the 5e-5 learning rate and then linearly decayed. To accommodate the long paragraphs of new data, we trained the model with a sequence length of 512. The overall training time for each Cont model took approximately 70 hours.

Fine-tuning Setup To perform the classification tasks, including binary classification and multi-label classification, we added a sequence classification/regression head on the pooled output of BERT. For the tasks, we utilized cross-entropy loss. Each dataset was divided into training, testing and development sets with proportions of (60,20,20).

To perform Named Entity Recognition task, we predicted the sequence of BIO tags, which is a common tagging format for tagging tokens in a chunking task, for each token in

Table 4.1. Summary of selected BERT-base models in general and specific domains.

Model	Domain	Corpora	Size
BERT	General	Wiki + Books	3.3B words/16GB
BioBERT	Bio-medical	PubMed	4.5B words
SciBERT	Science	BIO + CS papers	3.2B words
BlueBERT	Bio-medical	PubMed + MIMIC	4.5B words
PubBERT	Bio-medical	PubMed	3.2B words/21GB
LegalBERT	Law	legislation + court cases	12GB
ConfliBERT	Conflict	organization/government reports + news	34GB

the input sentence. We pre-processed the dataset to ensure that the input has the correct CoNLL format (Sang and De Meulder, 2003). Each dataset was divided into training, testing and development sets with proportions of (60,20,20).

We conducted fine-tuning of our models on a single Nvidia A-100 GPU, iterating over 5 epochs. The learning rate was 5e-05, and a batch size was 16. For Named Entity Recognition, the maximum sequence length was 128, while for Classifications, it was 512. To ensure the robustness, we repeated all experiments 10 times with distinct seeds. The performance of the models was evaluated using F1 scores, which served as the performance metric for all the tasks.

4.2 Pre-training

The pre-training of domain-specific BERT is the process of further training an existing model on a domain-specific corpus to adapt it to a specific domain. In our work, pre-training is to further train pre-trained Multilingual BERT and BETO on a conflict domain-specific corpus. To pre-train, we need to build a conflict domain-specific corpus and pre-train our model on it. As we follow the standard method for pre-training BERT-based model (https://github.com/huggingface/transformers), we will focus on explaining how we built the conflict domain-specific corpus in the rest of this section.

We utilized three types of datasets for pre-training ConfliBERT Spanish. For the first type, we used texts that are crawled from highly conflict related categories, which are ”politic” and ”international”, of Spanish news websites. The list of Spanish news websites are listed in Table 4.2. Secondly, we crawled texts from NGO websites articles that uses Spanish language. Since NGO websites deal with various conflicts in society, purely conflict related texts can be obtained. The list of NGO websites are listed in Table 4.3. Lastly, we utilized two open dataset that contains conflict related Spanish corpus. They are listed in Table 4.4.

4.2.1 Data Acquisition

News websites

We obtained Spanish texts related to political conflict and violence from news websites. We utilized 123 news websites from 18 Spanish using countries. We mainly crawled text from news articles in politically related categories from each websites. We selected the categories based on the assumption that these categories would have a higher concentration of news related to conflicts. We excluded articles from the not political categories, such as economy, business and sports. In total, the dataset we collected from news websites is approximately 7.8 GB in size. The list of the news websites are listed in Table 4.2.

NGO websites

We crawled the relevant texts from NGO websites that are written in Spanish. Among our research team, political scientists selected NGOs dealing with political conflicts, and the articles on these sites were crawled on the assumption that they were highly related to conflicts. The dataset we collected from 97 NGO websites from 8 Spanish using countries is approximately 1.1 GB in size. The list of the NGO websites are listed in Table 4.3.

Public dataset

We utilized Spanish texts related to political conflict and violence that is obtained from two open access datasets, which are MultiUN (Eisele and Chen, 2010) and DGT (Tiedemann, 2012). The MultiUN dataset is a multilingual corpus that consists of official United Nations documents translated into multiple languages. The DGT dataset is a multilingual corpus created by the European Union's Directorate-General for Translation. We used Spanish part of theses corpus. The dataset we obtained from the open access datasets is approximately 2.8 GB in size. The public websites are listed in Table 4.4.

In total, we built political conflict related Spanish corpus in size near 11.7 GB. For the computing resource, we used High Performance Computing (HPC) resource of University of Arizona (UA). We built Python scripts for crawling, and utilized Python packages, such as Newspaper, BeautifulSoup and Requests, to extract text from websites.

Table 4.2. List of data acquisition sources-News websites

Country	News Website	Size
Argentina	Alai	133,371 KB
	Ambito	643 KB
	Clarin + CS papers	237,244 KB
	Diario San Rafael	304 KB
	Diario Hoy	173,685 KB
	Diario El Argentino	1,052 KB
	El Comercial	10,712 KB
	El Cordillerano	1,544 KB
	El Independiente	28,092 KB
	Surenio	41,289 KB
	La Arena	19,731 KB
	La Semana	3,331 KB
	Lavoz	69,580 KB
	Los Andes	68,026 KB
	Diario Rio Negro	54,760 KB
Aruba	Diario Online	30,638 KB
Bolivia	ABI	10,636 KB
	El Diario - Bolivia	14,903 KB
	El Mundo - Bolivia	3,808 KB
	El Pais Tarija	392,976 KB
	Jornada	11,557 KB
	La Razon	257,653 KB
	Los Tiempos	2,115 KB
	Opinion	167,217 KB
Chile	El Ciudadano	24,409 KB
	El Mostrador	32,147 KB
	La Nacion	15,486 KB
Colombia	El Diario - Colombia	3,782 KB
	El Nuevo Dia	116,362 KB
	El Nuevo Siglo	64,932 KB
	El Tiempo	8,900 KB
Dominican Republic	Diario Libre	31,500 KB
	El Caribe	10,100 KB
	El Nuevo Diario	47,500 KB
Ecuador	El Heraldo	2,800 KB
	El Mercurio	21,700 KB
	El Telegrafo	60,000 KB
	Diario Los Andes	808 KB

Table 4.2 continued

Country	News Website	Size
	La Primicia	200 KB
	Machala Movil	242 KB
Guatemala	Prensa Libre	37,100 KB
	El Periodico - Guatemala	2,100 KB
	El Metropolitano	7,200 KB
	Republica - Guatemala	10,500 KB
	Aldia - Guatemala	6,000 KB
Honduras	Primicia Honduras	7,000 KB
	STN Honduras	5,400 KB
	Diario QuienOpina	49,200 KB
	Diario Paradigma	59,200 KB
	El Mundo - Honduras	1,650 KB
	En Alta Voz - Guatemala	3,600 KB
Mexico	El Heraldo de Aguascalientes	4,600 KB
	La Voz De Michoacan	30,400 KB
	El Sol De Morelia	293 KB
	Cambio De Michoacan	6,400 KB
	Quadratin	1,100 KB
	El Sol De Mexico	596 KB
	El Sol De Centro	368 KB
	El Vigia	2,000 KB
	El Heraldo De Chihuahua	375 KB
	Cuarto Poder	111 KB
	Tribuna	2,300 KB
	El Solde Puebla	550 KB
	24 Horas	65,900 KB
	La Razon De Mexico	43,000 KB
	La Prensa - Mexico	508 KB
	Capital Mexico	1,140 KB
	Diario de Xalapa	640 KB
	El Sol De Zacatecas	1,000 KB
Nicaragua	Confidencial	49,100 KB
	La Jornada	237 KB
	La Prensa Nicaragua	431,700 KB
	Articulo 66	5,400 KB
	Nicaragua Actual	21,400 KB

Table 4.2 continued

Country	News Website	Size
Panama	Dia a Dia	18,388 KB
	Panama America	4,320 KB
	Critica	5,640 KB
Paraguay	La Nacion - Paraguay	1,190 KB
Peru	La Razon	13,901 KB
	Diario Expreso	156,099 KB
	Enlinea.pe	9,912 KB
Uruguay	Semanario Cronicas	12,306 KB
	Red Del Tercer Mundo	157 KB
	El Pais 24	1,545 KB
Venezuela	Correo Del Orinoco	9,885 KB
	El Impulso	126,866 KB
	El Periodiquito	5,516 KB
	Diario Veo	6,571 KB
	La Patilla	443,626 KB
	El Pitazo	71,165 KB
Spain	El Periodico Extremadura	344,000 KB
	El Progreso	130,000 KB
	Noticias De Gipuzkoa	341,000 KB
	Dia De Ibiza	63,000 KB
	El Periodico Mediterraneo	354,000 KB
	La Opinion A Coruna	99,377 KB
	Eldia.es	158,972 KB
	Diario Cordoba	52,840 KB
	Le Region	211,114 KB
	Granada Hoy	420,128 KB
	Malaga Hoy	729,000 KB
	ABC.es	23,489 KB
	El Mundo	35,425 KB
	Independent en Espanol	240,460 KB
	El Periodico	732,585 KB
	El Correo	584,599 KB
	El Diario Vasco	679,093 KB
	Diario De Navarra	211,816 KB
	La Provincias	508,988 KB
	ABC De Sevilla	669,694 KB

Table 4.2 continued

Country	News Website	Size
International	UN News	75,478 KB
	BBC News Mundo	14,158 KB
	La Semana	3,227 KB
	Agencia EFE	98,761 KB
	La Semana	3,227 KB
	Eurones Spanish	141,400 KB
	Latino Rebels	52,901 KB
	Latin America News Dispatch	7,417 KB
	Univision	52,901 KB
	The Conversation	5,600 KB

Table 4.3. List of data acquisition sources-NGOs

Country	NGO	Size
International	Organización de Estados Americanos	32,130 KB
	Corte Interamericana de Derechos Humanos	266,000 KB
	Alto Comisionado de las Naciones Unidas para los Derechos Humanos	38,200 KB
	Human Rights Watch	34,400 KB
	Amnistia Internacional	48,400 KB
	Comisión Interamericana de Derechos Humanos	25,300 KB
	Médicos sin Fronteras	12,550 KB
	Cruz Roja	55,800 KB
	Instituto Interamericano de Derechos Humanos	162,000 KB
	Federación Iberoamericana Ombudsman	8,090 KB
	Federación Internacional por los Derechos Humanos	2,420 KB
	Organización Mundial Contra la Tortura	2,990 KB
	La Red de Instituciones Nacionales de Derechos Humanos	2,440 KB
	Derechos Digitales	2,110 KB
	ONU Mujeres	3,630 KB
	ACNUR	12,500 KB
	Comittee to Protect Journalists	12,100 KB
	Comité por los Derechos Humanos en América Latina	3,330 KB
	Iniciativa Mesoamericana de Mujeres Defensoras de Derechos Humanos	2,790 KB
	Protection Internacional	453 KB
	Agenda Estado de Derecho	1,460 KB
	WOLA	9,470 KB
	Centro de Estudios de Justicia de las Américas	102 KB
Argentina	Amnistía Internacional Argentina	50,800 KB
	Asamblea Permanente por los Derechos Humanos	1,800 KB
	Asociación de Ex-Detenidos Desaparecidos	949 KB

Table 4.3 continued

Country	NGO	Size
	Asociación de Madres de la Plaza de Mayo	5,041 KB
	Asociación Civil por la Igualdad y la Justicia	5,990 KB
	Centro de Estudios Legales y Sociales	70,600 KB
	Centro de Profesionales por los Derechos Humanos	43,400 KB
	Coordinadora Contra la Represión Policial e Institucional	1,870 KB
	Equipo Argentino de Antropología Forense	532 KB
	Familiares de desaparecidos y detenidos por razones políticas de Córdoba	1,520 KB
	Hijos por la Identidad y la Justicia contra el Olvido y el Silencio	765 KB
	Memoria Abierta	866 KB
	Arte y Esperanza Asociación Civil	249 KB
	Comité de Acción Jurídica	315 KB
	Colectivo al Margen	1,360 KB
	Centro para la Apertura y el Desarrollo de América Latina	16,800 KB
	Comisión Argentina para Migrantes y Refugiados	3,300 KB
Bolivia	Centro de Estudios Jurídicos e Investigación Social	13,316 KB
	Amnistía Internacional Bolivia	490,000 KB
	Fundación Solón	2,080 KB
	Católicas por el Derecho a Decidir	779 KB
	ADESPROC LIBERTAD GLBT	1,550 KB
	Oficina Jurídica para la Mujer	274 KB
	Observatorio de los derechos LGBT	32,500 KB
	IPAS Bolivia	2,530 KB
	Fundación Tribuna Constitucional Plurinacional Bolivia	1,370 KB
	Internet Bolivia	250 KB
	The Conversation	5,600 KB
Chile	Amnistía Internacional	2,040 KB
	Corporación para Comunidad y Justicia	1,970 KB
	Corporación de Promoción y Defensa de los Derechos del Pueblo	716 KB
	Centro de Estudios de la Realidad Social	1,090 KB
	Todo Mejorar	130 KB
	Organización Trans Diversidades	2,476 KB
	Fundación Iguales	936 KB
	Movimiento por la Diversidad Sexual MUMS	18,370 KB
	Fundación de Documentación y Archivo de la Vicaría de la Solidaridad	554 KB
	Museo de la Memoria y los Derechos Humanos	3,580 KB
	Observatorio contra el Acoso Chile	213 KB
	Fundación Instituto de la Mujer	211 KB
	Rompiendo el Silencio	433 KB
	Fundación Instituto Indígena	263 KB
	Asociación por la Memoria y los Derechos Humanos Colonia Dignidad	330 KB
	Corporación de Memoria y Cultura de Puchuncaví	169 KB
	Centro Cultural Museo y Memoria de Neltume (CCMMN)	189 KB

Table 4.3 continued

Country	NGO	Size
	Agrupación de Familiares de Ejecutados Políticos	2,570 KB
	Red Internacional de Apoyo a los Presos Políticos de Chile	113 KB
	Fundación de Protección a la Infancia Dañada por los Estados de Emergencia	3,980 KB
Colombia	Instituto Latinoamericano para una Sociedad y un Derecho Alternativos	103 KB
Costa Rica	Centro Cultural Museo y Memoria de Neltume (CCMMN)	189 KB
	Asociación Universal de Embajadores para la Paz	3,630 KB
	Fundación Justicia y Género	584 KB
	Fundación CEPPA	102 KB
	Fundación Acceso	114 KB
	Organización Internacional para las Migraciones (OIM)	477 KB
	Facultad Latinoamericana de Ciencias Sociales	66 KB
Dominican	Alianza ONG	911 KB
	Fundación Solidaridad	780 KB
	Participación Ciudadana	5,830 KB
	Amnistía Internacional República Dominicana	33,900 KB
	Museo Memorial de la Resistencia Dominicana	40 KB
Ecuador	INREDH, por los derechos humanos, de los pueblos y de la naturaleza	5,670 KB
	Comisión Ecuménica de Derechos Humanos	641 KB
	Surkunaa	36 KB
	Amazon Frontlines	1360 KB
	Comité Permanente por la Defensa de los Derechos Humanos	501 KB
	Fundamedios	60 KB
	Coalición Nacional de Mujeres del Ecuador	152 KB
	ACDemocracia	38 KB
	Asociación Latinoamericana para el Desarrollo Alternativo, ALDEA	528 KB
	Fundación Alejandro Labaka (FAL)	256 KB
	Paz y Desarrollo	328 KB
	Colectivo Geografía Crítica de Ecuador	741 KB
	Instituto de Estudios Ecuatorianos	293 KB

Table 4.4. List of data acquisition sources-others

Dataset	Source	Size
MultiUN	Collection of translated documents from United Nations	2,140,000 KB
DGT	Collection of translated documents from European Union's Directorate-General for Translation	687,000 KB

4.2.2 Data Pre-processing

Data Cleaning We used the existing Huggingface code for pre-training ConfliBERT Spanish. The code includes most of basic text pre-processing steps. Some general pre-processing steps are unnecessary due to the nature of the BERT based models. Therefore, typical text pre-processing steps, such as stop word removal and lemmatization are not required for our task. We only removed peripheral punctuation marks and extra white space before and after text which are crawled not intentionally from advertisements or unique page structure.

Data Filtering Even though we crawled text only from politically related categories, we could still observe irrelevant texts in the crawled corpus. For example, there were the texts, crawled from international category of news website, talking about an international events such as Olympic matches. The texts are irrelevant to political conflict and violence, so those can harm the validity of our model. Therefore, we built a filter based on the keywords provided by political scientists that can filter out irrelevant texts from the corpus.

We utilized two types of keywords, relevant and irrelevant keywords, which were created after verbs and actors in the CAMEO dataset (Parolin et al., 2022). Then, we augmented and revised the keyword list with the aid of political scientist to make it larger and more accurate. Most of the observed example that we intended to filter out came from sports articles. The international category contains quite amount of sports articles, and they are barely detected only by related keywords because they contain keywords like "attack" which can be seen as conflict relevant. Thus, we added irrelevant keywords list to exclude not related texts from corpus. We then compared the number of matches between the relevant and irrelevant keywords and tuned the thresholds with the assistance of experts to extract the most appropriate conflict related news. The relevant and irrelevant keywords are described in Table 4.5. and Table 4.6. In the tables, we provided Spanish keyword with English-translated keyword to help understanding.

Table 4.5. List of relevant keywords.

Spanish	English	Spanish	English
abuso	abuse	ciberseguridad	cybersecurity
activista	activist	civil	civil
actos	acts	coaccionar	coerce
administración	administration	colgado	strung up
agencias	agencies	colonial	colonial
alcalde	mayor	combate	combat
allegati	attachments	comité	committee
amenaza	threat	comunismo	communism
anarquía	anarchy	comunista	communist
aplicación	application	concejal	councilor
apuñalado	stabbed	condenado	condemned
apuñalamiento	stabbing	conflicto	conflict
arma	weapon	congreso	congress
armado	armed	conscripto	conscript
armarios	wardrobes	consejo	advice
armas	weapons	conservador	conservative
artillería	artillery	constitución	Constitution
asalto	assault	constituenc	constituency
asamblea	assembly	contra el terrorismo	against terrorism
asesinato	murder	contra las mujeres	against women
asesino	killer	contrainsurgencia	counterinsurgency
asilo	asylum	convicto	convicted
asuntos exteriores	foreign affairs	corps	corps
asuntos sociales	social issues	corrupto	corrupt
ataque	stroke	criminal	criminal
attrocit	attrocit	cuerpos	bodies
autoridades	authorities	cárceles	prisons
bajas	low	daño	damage
batallas	battles	defensa	defending
bienestar	welfare	delegado	delegate
blotter	blotter	delitos	crimes
boicot	boycott	democrático	democratic
boleta electoral	electoral ticket	demostración	demonstration
bomba	bomb	departamento	department
brexit	brexit	deportar	deport
casa blanca	White House	derechos	rights
caso	case	derramamiento de sangre	bloodshed
casualt	casualty	desarmado	disarmed
censores	censors	desestablecido	disestablished
ciberataque	cyber attack	desigualdad	inequality
cibercrimen	cybercrime	desobedecer	disobey

Table 4.5 continued

Spanish	English	Spanish	English
desobediencia	disobedience	forcibl	forceful
desplazar	displace	formaciones	formations
desplegar	deploy	fosa masiva	mass grave
destruir	destroy	fraude	fraud
detención	detention	frontera	border
detener	arrest	fuerzas	forces
dictador	dictator	funcionarios	civil servants
diplomático	diplomatic	genocidio	genocide
diputado	diputado	gobernantes	rulers
discriminar	discriminar	gobernar	govern
dispara	shoot	golpes de estado	hit of State
disparando	shooting up	granada	grenade
disparo	Shooting	guerra	war
disputa	quarrel	guerras	wars
disputas	disputas	guerrilla	warfare
disturbios	unrest	gángster	gangster
drogas	drugs	huelga	strike
ejecuciones	executions	ilegal	illegal
ejecutado	executed	incendiario	incendiary
ejecutar	execute	incidentes	incidents
ejército	army	independencia	independence
elecciones	elections	injur	injury
electoral	electoral	inmigración	immigration
emancipat	emancipate	insurgente	insurgent
embajada	embassy	inteligencia	intelligence
embajador	ambassador	intergubernamental	intergovernmental
encarcelar	imprison	invadir	encroach
esclavizar	enslave	invasión	invasion
esclavo	slave	jueces	judges
estados miembros de	member states of	juicios	judgments
expatriados	expats	justicia	justice
explosión	burst	legalidad	legality
explotar	blow	legalización	legalization
expulsar	expel	legisladores	legislators
extraditar	extradite	legislativo	legislative
extranjero	foreign	leyes	laws
extremista	extremist	liberación	release
federal	federal	liberal	liberal
feminista	feminist	lobby	lobby

Table 4.5 continued

Spanish	**English**	**Spanish**	**English**
mancomunidad	commonwealth	prensa libertad	press freedom
mantenimiento de la paz	peace keeping	preprisal	preprisal
marihuana	dope	presidente	president
masacre	slaughter	prisión	prison
matanza	slaughter	propaganda	propaganda
matar	kill	prostitut	prostitute
medios de comunicación de masas	mass media	protesta	protest
milicia	militia	rebelde	rebel
militante	militant	referéndum	referendum
militar	military	reforma	reform
ministerio	ministry	refugio	shelter
ministerios	ministries	relaciones	relations
ministro	minister	relaciones	relations
misiles	missiles	relaciones exteriores	external relationships
monarca	monarch	relación internacional	international relation
movimiento	motion	representación	representation
muerte	death	reprimir	suppress
municipal	municipal	republicano	republican
mutilación	mutilation	resolución	resolution
nacionalidad	nationality	restringir	restrict
nacionalismo	nationalism	rifle	rifle
nacionalista	nationalist	sanción	sanction
NATO	NATO	secretario de estado	secretary of state
non violen	non violent	secuestrar	kidnap
NYPD	NYPD	seguridad	security
ocupación	occupation	senado	senate
oficial	official	separatismo	separatism
operaciones	operations	servidores	servant
organización	organization	socialista	socialist
organizado	organized	soldado	soldier
parlamento	parliament	sospechoso	suspicious
partido	game	supremo	supreme
partidos	match	territorial	territorial
países en	countries in	territorio	territory
persecución	persecution	terror	terror
piquete	picket	think tank	think tank
policía	police	titular de la oficina	head office
política	policy	tomar represalias	retaliate
político	political	tortura	torture

Table 4.5 continued

Spanish	English	Spanish	English
tragedia	tragedy	vigilancia	surveillance
tratados	treaties	violación	rape
tribunal	court	violat	violated
tribunales	courts	violen	violate
tropa	troop	voto	vote
tráfico	traffic	víctima	victim
trágico	tragic		

Table 4.6. List of irrelevant keywords.

Spanish	English	Spanish	English
accidente	accident	bádminton	badminton
afiliación	membership	campeonato	championship
agricultura	agriculture	campeón	champion
alimentos	food	campo	field
amex	amex	cantante	singer
animal	animal	cardenales	cardinals
arte	art	carrera	career
artista	artist	carros	cars
athlet	athlete	cartelera	billboard
australian open	australian open	cba playoffs	cba playoffs
auto rac	car racing	celebrid	celebrity
automóviles	automobiles	cerveceros	brewers
autos	cars	cine	cinema
aventura	adventure	clima	climate
bbc three	bbc three	clinics	clinics
bcs campeón	bcs champion	coches	cars
bcs nacional	national bcs	comed	eat
belleza	beauty	comercio	trade
bestilos	bestilos	comidas	foods
billete	ticket	commentisfree	commentisfree
bola	ball	commodit	commodity
bolos	bowling	compradores	buyers
boxeo	boxing	compras	shopping
braves	braves	concierto	concert
british open	british open	consumidor	consumer
broadway	broadway	copa	cup
corporat	corporate	goles	goals

Table 4.6 continued

Spanish	English	Spanish	English
cotización	price	golf	golf
cricket	cricket	grammy	grammy
cultura	culture	grand national	grand national
cultura popular	popular culture	grand prix	grand prix
cyclyng	cycling	grand slam	grand slam
daytona 500	daytona 500	gráfico	graphic
deporte del motor	motor sport	género	gender
deportes	sports	hipoteca	mortgage
desastre natural	natural disaster	hockey	hockey
dinero	money	hollywood	hollywood
disparar	shoot	huracán	hurricane
djia	djia	indianápolis 500	indianapolis 500
dow jones	dow jones	indycar	indycar
dragster	dragster	info	info
dólar	dollar	inmobiliario	real estate
economía	economy	intercambio	exchange
ecuestre	equestrian	invertir	invest
educación	education	iron man	iron man
entrenamiento de primavera	spring training	jazz	jazz
entretenimiento	entertainment	juegos	games
especiales	specials	jugador	player
estilo de vida	Lifestyle	jugar bola	play ball
expos	expos	kentucky derby	kentucky derby
ficción	fiction	lanzadores	pitchers
fifa	fifa	le mans	le mans
fina	fine	libros	books
finales	finals	libros revisar	book review
finalista	finalist	liga	league
financ	finance	locura de marzo	march madness
financiación	financing	lotería	lottery
fitness	fitness	lpga	lpga
flowery gold minesflod	flowery gold minesflod	maratón	marathon
fotos	photos	mariscal de campo	Quarterback
francés abierto	french open	meast	meast
fuera de temporada	out of season	medallista	medalist
fund	fund	medio ambiente	environment
futuro	future	medio tiempo	halftime
fútbol	soccer	mercado	market
galería	Gallery	mets	mets
ganancia	revenue	miembros	members
ganar	gain	moda	fashion
gimnástico	gymnastic	moneda	currency
goleadores	scorers	mundo del espectáculo	showbiz

Table 4.6 continued

Spanish	**English**	**Spanish**	**English**
música	music	recapitulación	recapitulation
nadar	swim	receta	recipe
nascar	nascar	regalos	gifts
nasd	nasd	religión	religion
nasdaq	nasdaq	remo	rowing
nba	nba	rendimiento	performance
ncaa	ncaa	resultados	results
negocio	business	resumen	summary
netflix	netflix	rugby	rugby
nfl	nfl	russell 2000	russell 2000
nhl	nhl	russell us index	russell us index
novelas	novels	ryder	ryder
nyse	nyse	salud	health
obituar	obituate	seis naciones	six nations
ofertas	offers	semi final	semi final
ofertas bcs	bcs offers	stock	stock
olímpico	olympic	surf	surf
orioles	orioles	surfista	surfer
oscar	oscar	t-bill	t-bill
padres	parents	teatro	theater
paidpost	paidpost	tech	technology
paquetes	packages	tecnología	technology
paralímpico	paralympic	temblor de corazón	tremor of heart
parrilla de salida	grille output	temperatura	temperature
película	movie	tenis	tennis
películas	films	tenía meta	had goal
peso leventar	weight lift	tesorería	treasury
pga	pga	tesoros	treasures
phillies	phillies	tiempo	time
pingüinos	penguins	tierra	land
piratas	pirates	tipo de interés	type of interest
playoffs	playoffs	tiro	shot
podcast	podcast	top sing	top sing
pollut	polluted	tornado	tornado
polo	pole	torneo	tournament
pop	pop	tour de francia	tour de France
popular	popular	triatlón	triathlon
portero	goalkeeper	turismo	tourism
powerball	powerball	tv	tv
precios	prices	us open	us open
producto	product	uefa	uefa
pse	pse	vela	candle
puntuación	punctuation	vendedor	seller
radio	radio	ventas	sales

Table 4.6 continued

Spanish	English	Spanish	English
viaje	journey	vida	life
videos	videos	vr	vr
wimbledon	wimbledon	world classic	world classic
world series	world series	yankees	yankees
álbum	album	éxito	success
índice futuro	future index		

4.3 Fine-tuning

The ConfliBERT Spanish is trained, and the model needs to be fine-tuned on specific tasks to be applied to each downstream tasks. We follow the standard BERT fine-tuning process (https://github.com/huggingface/transformers) as described in Figure 3.6.. We focus on explaining dataset that we used for fine-tuning in the following part of this section.

4.3.1 Huffingtonpost

We crawled text from Huffingtonpost Spanish website to use it for binary classification task. We collected politically relevant text from news articles in politics and international categories, and politically irrelevant text from those in sports and economy categories. We labeled relevant text and irrelevant text as "0" and "1", respectively. The dataset contains 3130 rows in total and the number of both labels are 2019 and 1111.

4.3.2 Protest

We built Protest dataset to evaluate models on binary classification and multi-label classification tasks. The dataset contains annotations of 723 Spanish news articles related to social protest from Associated French Press (AFP) extracted from Gigaword Spanish (Mendonça et al., 2006). The corpus is a random selection of news articles including the word "protest" from 1994 to 2006. The annotations contain 4 classes on action, which are material conflict, verbal conflict, material cooperation, and verbal cooperation. For binary classification task, we labeled actions that contain conflict, which are material conflict and verbal conflict, as "0" and those that contain cooperation, which are material cooperation and verbal cooperation as "1". The Protest dataset for binary classification contains 723 rows in total and the number of labels for conflict and cooperation are 411 and 312. For multi-label classification task, we labeled each labels "0" if original text contains the corresponding annotation. If not, we labeled "1". The Protest dataset for multi-label classification contains 723 rows in total and the number of material conflict, verbal conflict, material cooperation, and verbal cooperation labels are 337, 440, 663, and 585, respectively.

4.3.3 InsightCrime

We used InsightCrime dataset (Parolin et al., 2021) on model evaluation on multi-label classification tasks. The dataset consists of news articles reporting organized crime activity in both English and Spanish. The corpora came from the InsightCrime web page (https://www.insightcrime.org). We only used Spanish part of the InsightCrime dataset. The dataset contains 22 types of action. We chose 4 actions, which are law enforcement, drug trafficking, homicides, and corruption, that appear more than 500 times since other actions appeared very little. The InsightCrime dataset for multi-label classification contains 2084 rows in total and the number of label for law enforcement, drug trafficking, homicides, and corruption are 672, 521, 368, and 268, respectively.

4.3.4 Mx-News

We used Mx-news dataset to evaluate models on named entity recognition tasks (Ramos-Flores et al., 2020). The dataset was built on the political news using 250 documents. It is in the Spanish language and it has seventeen classes for entities, which are "PER", "ORG",

Table 4.7. Details of labels in Mx-news dataset.

No.	Class	Description	Count
1	PER	People names, aliases and abbreviations	6,863
2	ORG	Organizations, institutions	4,779
3	DAT	Dates on different formats	4,530
4	TIT	Title or position of persons	3,696
5	GPE	Country names, states, cities, municipalities	2,201
6	PEX	Political party names, aliases and abbreviations	1,263
7	TIM	Time expresions	1,206
8	FAC	Facility names	821
9	EVT	Event names	802
10	ADD	Addresses expressions, URLs and Twitter users	740
11	MNY	Monetary amounts	715
12	DOC	Documents, laws, rules	669
13	PRO	Product names, brands, application names	506
14	PRC	Percentage expressions	338
15	DEM	Geographical or racial origin of people	294
16	AGE	People age	177
17	LOC	Locations about regions, rivers, lakes	131

"DAT", "TIT", "GPE", "PEX", "TIM", "FAC", "EVT", "ADD", "MNY", "DOC", "PRO", "PRC", "DEM", "AGE", and "LOC". More details about the classes are explained in Table 4.7. We formatted the dataset to CoNLL format (Sang and De Meulder, 2003).

4.4 Results

We used F1 scores to evaluate the performance of each ConfliBERT Spanish model on the downstream tasks. The scores are reported in Table 4.6.. In the table, Multilingual BERT is described as mBERT for convenience. Also, task types are abbreviated for convenience, binary classification to BC, multi-label classification to MLC, and named entity recognition to NER. For each tasks, we applied for every versions of ConfliBERT Spanish, which are built on top of Multilingual BERT cased, Multilingual BERT uncased, BETO cased, and BETO uncased, respectively. Also, baseline models, which are Multilingual BERT cased, Multilingual BERT uncased, BETO cased, and BETO uncased, were fine-tuned on each tasks.

As indicated in Table 4.8., ConfliBERT Spanish continuously outperformed Multilingual BERT and BETO baselines across all the tasks. The model showed best result in each task are highlighted in bold, and our models consistently achieved the best results. The results demonstrate the superiority of ConfliBERT Spanish achieving better scores in every case, regardless of whether it is Multilingual BERT or BETO, cased or uncased.

Binary Classification First, we used Huffingtonpost dataset that contains Spanish news articles crawled from politics category for relevant text, and economy and sports category for irrelevant text. We set label as "political" for and "non-political". The models were fine-tuned to classify whether given text is related to politics or not.

Next, we used Protest dataset that contains Spanish news articles related to social protest extracted from Gigaword Spanish (Mendonça et al., 2006). We set label conflict text as "conflict" and cooperation text as "non-conflict". The models were fine-tuned to classify whether given text contains conflict or not.

Our models showed improved performance across both binary classification tasks in all cases. Therefore, we can say that ConfliBERT Spanish showed excellence in the binary classification for Spanish text that contains political conflict and violence.

Multi-label Classification First, we used InsightCrime dataset (Parolin et al., 2021) that contains Spanish news articles reporting organized crime activities. We chose the 4 most frequent labels among original 17 labels, since most of the labels had very low frequency. The models were fine-tuned to classify whether given text is relevant to 4 labels, which are "Law Enforcement", "Drug Trafficking", "Homicides", and "Corruption".

Next, we used Protest dataset that contains Spanish news articles related to social protest extracted from Gigaword Spanish (Mendonça et al., 2006) as mentioned. We assigned "Material Conflict", "Verbal Conflict", "Material Cooperation", and "Verbal Cooperation" labels individually if the original data contains the regarding annotation. Due to the nature of the multi-label classification task, each task can be assigned to more than one label. The models were fine-tuned to classify which labels a given text has.

Our models showed improved performance across both multi-label classification tasks in all cases. Therefore, we can say that ConfliBERT Spanish showed excellence in the multi-label classification for Spanish text that contains political conflict and violence.

Named Entity Recognition For named entity recognition task, we used Mx-news dataset that was built on the political news domain using 250 documents. It is in the Spanish language and has seventeen classes. It is annotated using tagging schema IOBES. The models were fine-tuned on the dataset to recognize entity and figure out a type of the entity.

Our models showed improved performance. In all cases of the named entity recognition task, our model performed better than the baseline models. Therefore, we can say that ConfliBERT Spanish showed excellence in the named entity recognition task for Spanish text that contains political contents.

Table 4.8. Summary of F1 measure results for fine-tuned model evaluation

Dataset	Domain	Task	Models	mBERT		BETO	
				cased	uncased	cased	uncased
Huffingtonpost	Politics	BC	Baseline	0.8757	0.8629	0.8816	0.8750
			ConfliBERT Spanish	**0.8960**	0.8890	0.8897	0.8854
Protest	Conflict	BC	Baseline	0.7956	0.8364	0.8295	0.8554
			ConfliBERT Spanish	0.8401	0.8391	0.8296	**0.8725**
Insight Crime	Crime	MLC	Baseline	0.7449	0.7235	0.7578	0.7548
			ConfliBERT Spanish	**0.7774**	0.7713	0.7731	0.7615
Protest	Conflict	MLC	Baseline	0.5649	0.4688	0.5807	0.5810
			ConfliBERT Spanish	0.5799	**0.6348**	0.5973	0.5964
Mx News	Politics	NER	Baseline	0.8292	0.8269	0.8336	0.7872
			ConfliBERT Spanish	0.8327	0.8331	0.8360	**0.8396**

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this paper, we present ConfliBERT Spanish, a pre-trained language model specifically designed for analyzing political conflict and violence in the Spanish language. Developing ConfliBERT Spanish involved acquiring and curating a substantial corpus of domain-specific data for the pre-training phase. We conducted extensive evaluations of the model's performance on various NLP tasks and datasets, consistently demonstrating that ConfliBERT Spanish outperforms baselines, such as multilingual BERT and BETO, in the domain of conflict and political violence, especially when working with limited amount of data. These findings can be highly valuable to researchers and decision makers interested in monitoring, analyzing, and predicting political conflict and violence in the Spanish society.

In future research, there is potential to further investigate parameters such as vocabulary size and pre-training epochs that were not thoroughly analyzed in this study. It would be beneficial to optimize ConfliBERT in future work. Moreover, exploring the application of ConfliBERT to more complex tasks, including understanding, inference, question answering, and uncertainty qualification, would be of great interest.

REFERENCES

Abadji, J., P. O. Suarez, L. Romary, and B. Sagot (2022). Towards a cleaner document-oriented multilingual crawled corpus. *arXiv preprint arXiv:2201.06642*.

Abedin, M., S. Nessa, L. Khan, and B. Thuraisingham (2006). Detection and resolution of anomalies in firewall policy rules. In *Data and Applications Security XX: 20th Annual IFIP WG 11.3 Working Conference on Data and Applications Security, Sophia Antipolis, France, July 31-August 2, 2006. Proceedings 20*, pp. 15–29. Springer.

Abrol, S., V. Khadilkar, L. R. Khan, and B. M. Thuraisingham (2015, February 24). Systems and methods for determining user attribute values by mining user network data and information. US Patent 8,965,974.

Abrol, S. and L. Khan (2010a). Tweethood: Agglomerative clustering on fuzzy k-closest friends with variable depth for location mining. In *2010 IEEE Second International Conference on Social Computing*, pp. 153–160. IEEE.

Abrol, S. and L. Khan (2010b). Twinner: understanding news queries with geo-content using twitter. In *Proceedings of the 6th Workshop on Geographic information Retrieval*, pp. 1–8.

Al-Naami, K., S. Chandra, A. Mustafa, L. Khan, Z. Lin, K. Hamlen, and B. Thuraisingham (2016). Adaptive encrypted traffic fingerprinting with bi-directional dependence. In *Proceedings of the 32nd Annual Conference on Computer Security Applications*, pp. 177–188.

Alliance, O. E. D. (2015). Petrarch python engine for text resolution and related coding hierarchy.

Alsentzer, E., J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott (2019). Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Awad, M., L. Khan, F. Bastani, and I.-L. Yen (2004). An effective support vector machines (svms) performance using hierarchical clustering. In *16th IEEE international conference on tools with artificial intelligence*, pp. 663–667. IEEE.

Awad, M., L. Khan, and B. Thuraisingham (2008). Predicting www surfing using multiple evidence combination. *The VLDB Journal 17*, 401–417.

Awad, M. A. and L. R. Khan (2007). Web navigation prediction using multiple evidence combination and domain knowledge. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 37*(6), 1054–1062.

Ayoade, G., V. Karande, L. Khan, and K. Hamlen (2018). Decentralized iot data management using blockchain and trusted execution environment. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 15–22. IEEE.

Bagozzi, B. E., D. Berliner, and R. M. Welch (2021). The diversity of repression: Measuring state repressive repertoires with events data. *Journal of Peace Research 58*(5), 1126–1136.

Bahdanau, D., K. Cho, and Y. Bengio (2015). Neural machine translation by jointly learning to align and translate. In Y. Bengio and Y. LeCun (Eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Beger, A., C. L. Dorff, and M. D. Ward (2016). Irregular leadership changes in 2014: Forecasts using ensemble, split-population duration models. *International Journal of Forecasting 32*(1), 98–111.

Beieler, J. (2016). Generating politically-relevant event data. *arXiv preprint arXiv:1609.06239*.

Beltagy, I., K. Lo, and A. Cohan (2019). Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Bond, D., J. Bond, C. Oh, J. C. Jenkins, and C. L. Taylor (2003). Integrated data for events analysis (idea): An event typology for automated events data development. *Journal of Peace Research 40*(6), 733–745.

Brandt, P. T., V. D'Orazio, L. Khan, Y.-F. Li, J. Osorio, and M. Sianan (2022). Conflict forecasting with event data and spatio-temporal graph convolutional networks. *International Interactions 48*(4), 800–822.

Breen, C., L. Khan, and A. Ponnusamy (2002). Image classification using neural networks and ontologies. In *Proceedings. 13th International Workshop on Database and Expert Systems Applications*, pp. 98–102. IEEE.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. (2020a). Language models are few-shot learners. *Advances in neural information processing systems 33*, 1877–1901.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. (2020b). Language models are few-shot learners. *Advances in neural information processing systems 33*, 1877–1901.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez (2020). Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Chalkidis, I., M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos (2020). Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Chowdhary, K. and K. Chowdhary (2020). Natural language processing. *Fundamentals of artificial intelligence*, 603–649.

Clark, K., M. Luong, Q. V. Le, and C. D. Manning (2020). ELECTRA: pre-training text encoders as discriminators rather than generators. *CoRR abs/2003.10555*.

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

Eisele, A. and Y. Chen (2010). Multiun: A multilingual corpus from united nation documents. In *LREC*.

Glavaš, G., F. Nanni, and S. P. Ponzetto (2017). Cross-lingual classification of topics in political texts. Association for Computational Linguistics (ACL).

Golnabi, K., R. K. Min, L. Khan, and E. Al-Shaer (2006). Analysis of firewall policy rules using data mining techniques. In *2006 IEEE/IFIP Network Operations and Management Symposium NOMS 2006*, pp. 305–315. IEEE.

Gu, Y., R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH) 3*(1), 1–23.

Hamlen, K., M. Kantarcioglu, L. Khan, and B. Thuraisingham (2010). Security issues for cloud computing. *International Journal of Information Security and Privacy (IJISP) 4*(2), 36–48.

Hanna, A. (2017, Jan). Mpeds: Automating the generation of protest event data.

Haque, A., L. Khan, and M. Baron (2016). Sand: Semi-supervised adaptive novel class detection and classification over data stream. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 30.

Haque, A., L. Khan, M. Baron, B. Thuraisingham, and C. Aggarwal (2016). Efficient handling of concept drift and concept evolution over stream data. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pp. 481–492. IEEE.

Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. *Neural Computation 9*(8), 1735–1780.

Hu, Y., M. Hosseini, E. S. Parolin, J. Osorio, L. Khan, P. Brandt, and V. 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*, pp. 5469–5482.

Hussein, D. M. E.-D. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences 30*(4), 330–338.

Jacoby, T. (2007). *Understanding conflict and violence: Theoretical and interdisciplinary approaches*. Routledge.

Khan, L., M. Awad, and B. Thuraisingham (2007). A new intrusion detection system using support vector machines and hierarchical clustering. *The VLDB journal 16*, 507–521.

Khan, L. and D. McLeod (2000). Audio structuring and personalized retrieval using ontologies. In *Proceedings IEEE Advances in Digital Libraries 2000*, pp. 116–126. IEEE.

Kingma, D. P. and J. Ba (2015). Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun (Eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kowsari, K., K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown (2019). Text classification algorithms: A survey. *Information 10*(4), 150.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Latif, S., R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller (2020). Deep representation learning in speech processing: Challenges, recent advances, and future trends. *arXiv preprint arXiv:2001.00378*.

Lee, J., W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics 36*(4), 1234–1240.

Lewis, P., M. Ott, J. Du, and V. Stoyanov (2020). Pretrained language models for biomedical and clinical tasks: understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pp. 146–157.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lu, J. and J. Roy (2017). Universal petrarch: Language-agnostic political event coding using universal dependencies.

Luo, F., L. Khan, F. Bastani, I.-L. Yen, and J. Zhou (2004). A dynamically growing self-organizing tree (dgsot) for hierarchical clustering gene expression profiles. *Bioinformatics 20*(16), 2605–2617.

Luo, F., Y. Yang, J. Zhong, H. Gao, L. Khan, D. K. Thompson, and J. Zhou (2007). Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. *BMC bioinformatics 8*(1), 1–17.

Masud, M., J. Gao, L. Khan, J. Han, and B. M. Thuraisingham (2010). Classification and novel class detection in concept-drifting data streams under time constraints. *IEEE Transactions on knowledge and data engineering 23*(6), 859–874.

Masud, M. M., T. M. Al-Khateeb, L. Khan, C. Aggarwal, J. Gao, J. Han, and B. Thuraisingham (2011). Detecting recurring and novel classes in concept-drifting data streams. In *2011 IEEE 11th International Conference on Data Mining*, pp. 1176–1181. IEEE.

Masud, M. M., J. Gao, L. Khan, J. Han, and B. Thuraisingham (2008). A practical approach to classify evolving data streams: Training with limited amount of labeled data. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 929–934. IEEE.

Masud, M. M., L. Khan, and B. Thuraisingham (2007). A hybrid model to detect malicious executables. In *2007 IEEE International Conference on Communications*, pp. 1443–1448. IEEE.

Mendonça, , D. Jaquette, D. Graff, and D. DiPersio (2006). Spanish gigaword. In *Linguistic Data Consortium*.

Meta, A. (2023). Introducing llama: A foundational, 65-billion-parameter large language model. *Meta AI. https://ai. facebook. com/blog/large-language-model-llama-meta-ai*.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013b). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems 26*.

Nessa, S., M. Abedin, W. E. Wong, L. Khan, and Y. Qi (2008). Software fault localization using n-gram analysis. In *Wireless Algorithms, Systems, and Applications: Third International Conference, WASA 2008, Dallas, TX, USA, October 26-28, 2008. Proceedings 3*, pp. 548–559. Springer.

Norris, C., P. A. Schrodt, and J. Beieler (2017). Petrarch2: Another event coding program. *J. Open Source Softw. 2*(9), 133.

O'brien, S. P. (2010). Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International studies review 12*(1), 87–104.

Osman, A. S. (2019). Data mining techniques.

Osorio, J. and A. Reyes (2017). Supervised event coding from text written in spanish: Introducing eventus id. *Social Science Computer Review 35*(3), 406–416.

Osorio, J., A. Reyes, A. Beltrán, and A. Ahmadzai (2020). Supervised event coding from text written in arabic: Introducing hadath. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pp. 49–56.

Parolin, E. S., M. Hosseini, Y. Hu, L. Khan, J. Osorio, P. T. Brandt, and V. D'Orazio (2022). 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.*

Parolin, E. S., L. Khan, J. Osorio, P. T. Brandt, V. D'Orazio, and J. Holmes (2021). 3M-Transformers for Event Coding on Organized Crime Domain. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–10. IEEE.

Parolin, E. S., L. Khan, J. Osorio, V. D'Orazio, P. T. Brandt, and J. Holmes (2020). Hanke: Hierarchical attention networks for knowledge extraction in political science domain. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 410–419. IEEE.

Parveen, P., J. Evans, B. Thuraisingham, K. W. Hamlen, and L. Khan (2011). Insider threat detection using stream mining and graph mining. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pp. 1102–1110. IEEE.

Pennington, J., R. Socher, and C. D. Manning (2014a). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Pennington, J., R. Socher, and C. D. Manning (2014b). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Peters, M., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer (2018). Deep contextualized word representations. arxiv 2018. *arXiv preprint arXiv:1802.05365 12.*

Petrushin, V. A. and L. Khan (2007). *Multimedia data mining and knowledge discovery*, Volume 521. Springer.

Qiu, X., T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences 63*(10), 1872–1897.

Radford, A., K. Narasimhan, T. Salimans, I. Sutskever, et al. (2018). Improving language understanding by generative pre-training.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. (2019a). Language models are unsupervised multitask learners. *OpenAI blog 1*(8), 9.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. (2019b). Language models are unsupervised multitask learners. *OpenAI blog 1*(8), 9.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research 21*(1), 5485–5551.

Raleigh, C., r. Linke, H. Hegre, and J. Karlsen (2010). Introducing acled: An armed conflict location and event dataset. *Journal of peace research 47*(5), 651–660.

Ramos-Flores, O., D. Pinto, M. Montes-y Gómez, A. Vázquez, D. Pinto, V. Singh, and F. Perez (2020, jan). Probabilistic vs deep learning based approaches for narrow domain ner in spanish. *J. Intell. Fuzzy Syst. 39*(2), 2015–2025.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). Learning internal representations by error propagation.

Sahs, J. and L. Khan (2012). A machine learning approach to android malware detection. In *2012 European intelligence and security informatics conference*, pp. 141–147. IEEE.

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

Sanh, V., L. Debut, J. Chaumond, and T. Wolf (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Schrodt, P. A. and B. Hall (2006). Twenty years of the kansas event data system project. *The political methodologist 14*(1), 2–8.

Shaon, F., M. Kantarcioglu, Z. Lin, and L. Khan (2017). Sgx-bigmatrix: A practical encrypted data analytic framework with trusted processors. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1211–1228.

Sundberg, R. and E. Melander (2013). Introducing the ucdp georeferenced event dataset. *Journal of Peace Research 50*(4), 523–532.

Sutskever, I., O. Vinyals, and Q. V. Le (2014). Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, Volume 27. Curran Associates, Inc.

Thuraisingham, B., L. Khan, M. M. Masud, and K. W. Hamlen (2008). Data mining for security applications. In *2008 IEEE/IFIP International Conference on Embedded and Ubiquitous Computing*, Volume 2, pp. 585–589. IEEE.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Lrec*, Volume 2012, pp. 2214–2218. Citeseer.

Tu, M., P. Li, I.-L. Yen, B. M. Thuraisingham, and L. Khan (2008). Secure data objects replication in data grid. *IEEE Transactions on dependable and secure computing 7*(1), 50–64.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017a). Attention is all you need. *Advances in neural information processing systems 30*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin (2017b). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 30. Curran Associates, Inc.

Wang, L. and L. Khan (2006). Automatic image annotation and retrieval using weighted feature selection. *Multimedia Tools and Applications 29*, 55–71.

Wang, L., L. Liu, and L. Khan (2004). Automatic image annotation and retrieval using subspace clustering algorithm. In *Proceedings of the 2nd ACM international workshop on Multimedia databases*, pp. 100–108.

Ward, M. D., A. Beger, J. Cutler, M. Dickenson, C. Dorff, and B. Radford (2013). Comparing gdelt and icews event data. *Analysis 21*(1), 267–297.

Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems 32*.

Yen, I.-L., J. Goluguri, F. Bastani, L. Khan, and J. Linn (2002). A component-based approach for embedded software development. In *Proceedings Fifth IEEE International Symposium on Object-Oriented Real-Time Distributed Computing. ISIRC 2002*, pp. 402–410. IEEE.

Zhu, Y., R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pp. 19–27.

BIOGRAPHICAL SKETCH

Wooseong Yang completed his Bachelor of Science in Systems Management Engineering from Sungkyunkwan University, South Korea in February 2019. He continued to study in the same institution and completed his Master of Science in Industrial Engineering. During his graduate study in South Korea he studied at the Language Technologies Institute of Carnegie Mellon University for six months as a visiting scholar. After he graduated from the Sungkyunkwan University, he moved to Texas and started his Master of Science in Computer Science at The University of Texas at Dallas in Fall 2021. Currently, he is in his fifth and last semester of the program. After he finishes his fifth semester in Summer 2023, he is joining the Computer Science program at the University of Illinois at Chicago as a PhD student. His research interest is in overall data mining including natural language processing and recommender systems.

CURRICULUM VITAE

Wooseong Yang

August 2023

Contact Information:

Department of Computer Science
The University of Texas at Dallas
800 W. Campbell Rd.
Richardson, TX 75080-3021, U.S.A.

Email: wooseong.yang@utdallas.edu

Education:

BS, Systems Management Engineering, Sungkyunkwan University, 2019

MS, Industrial Engineering, Sungkyunkwan University, 2021
Source Domain Identification and Cross-domain Knowledge Transfer Method to Alleviate Data Sparsity Problem in Session-based Recommender Systems
Master's Thesis
Advisor: Mye Sohn

MS, Computer Science, The University of Texas at Dallas, 2023

Employment History:

Research Intern, Mycelebs, Inc, Summer 2016
Teaching Assistant, The University of Texas at Dallas, January 2023 – May 2023