MULTILINGUAL EXTRACTIVE QUESTION ANSWERING WITH CONFLIBERT FOR POLITICAL AND SOCIAL SCIENCE STUDIES

by

Parker M. Whitehead

APPROVED BY SUPERVISORY COMMITTEE:

Latifur Khan, Chair

Vibhav Gogate

Karen Mazidi

This thesis is dedicated to my family and friends

for being an endless well of love and support.

MULTILINGUAL EXTRACTIVE QUESTION ANSWERING WITH CONFLIBERT FOR POLITICAL AND SOCIAL SCIENCE STUDIES

by

PARKER M. WHITEHEAD, BS, MS

THESIS

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

MASTERS OF SCIENCE IN

COMPUTER SCIENCE

THE UNIVERSITY OF TEXAS AT DALLAS

August 2023

ACKNOWLEDGMENTS

I would like to express great appreciation to my advisor, Dr. Latifur Khan. In a world strife with violence, discrimination, and unrest, it is an honor for him to have given me the opportunity to work on ConfliBERT, a tool that could create genuine positive impact. His guidance, encouragement, and feedback have been instrumental to my research.

Additionally, I am incredibly grateful for Dr. Javier Osorio. His expertise in politics and endless support were both essential for my work, and I am truly lucky to have been able to be under his supervision.

I would like to express a deep gratitude to Dr. Karen Mazidi. At a crossroads of my time in academia, she went out of her way to encourage the pursuit of my dreams to one day become a professor of Computer Science. Without her, I would not be where I am today.

I would also like to acknowledge my appreciation for Dr. Vibhav Gogate. His lectures are some of the best I've had the pleasure of attending, and it is clear that he holds a very deep passion for machine learning. It was his teaching in CS 6347 that made me decide to focus in machine learning, which has been one of the best decisions of my life.

To my mother and father, thank you for being the best parents one could ask for. There have been countless times in which I have been lost, and you have found me. You are my rocks, and I am immensely fortunate to have you in my life.

Finally, I'd like to thank the many friends and colleagues I've had the pleasure of making throughout my time in academia. The culmination of their support, care, and encouragement have been the pillars that support my life. Truly, I could not do this without you all.

July 2023

MULTILINGUAL EXTRACTIVE QUESTION ANSWERING WITH CONFLIBERT FOR POLITICAL AND SOCIAL SCIENCE STUDIES

Parker M. Whitehead, MSCS
The University of Texas at Dallas, 2023

Supervising Professor: Latifur Khan, Chair

Political conflict and violence have emerged as prominent concerns for political scientists in both academia and policy circles. The overwhelming influx of complex and dense news makes it increasingly challenging to effectively monitor and analyze political events. To address this challenge and contribute to the advancement of conflict research, we propose the introduction of ConfliBERT English and ConfliBERT Spanish. These two domain-specific pre-trained language models are specifically designed for the analysis of political conflict and violence, and have undergone fine-tuning to excel in extractive question answering tasks, which are not susceptible to hallucination. The pre-training of our ConfliBERT models utilized our comprehensive conflict-specific corpus from diverse sources. In order to evaluate the performance of ConfliBERT for extractive question-answering, We performed fine-tuning on SQuAD v1.1 and NewsQA, two large question-answering datasets. Additionally, we created ConfliQA English and Spanish, two crowd-sourced evaluation datasets for conflict-domain extractive QA. Through extensive experimentation and evaluation on all versions of ConfliBERT English and Spanish, we proved that ConfliBERT English outperforms in analyzing political texts compared to BERT English baseline models, and provided detailed insight into further developing ConfliBERT for low-resource languages.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Conflict and violence is ever evolving throughout history, taking on new shapes and contexts. However, it's detrimental impact on humanity as a whole is a constant, leading to the loss of countless lives and tense division between government bodies (United Nations, 2020). A subset of political science known as conflict research (Jacoby, 2007) focuses on the analysis and comprehension of complex political subjects, including as war, terrorism, human rights, and more. Through a heightened understanding, researchers aim to apply their findings to the predicting and prevention of political conflict.

To enhance the effectiveness of conflict study, researchers have increasingly turned to the utilization of technology for political research. Initially, computers were employed for manual coding (Raleigh et al., 2010) to analyze specific political conflicts or events. However, the time-consuming process of developing competent programs and the limited applicability to other conflicts (Sundberg and Melander, 2013) rendered these approaches largely ineffective and challenging to maintain. Consequently, there was a shift towards data mining-based automated systems 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).

Data mining involves extracting information from large datasets in order to uncover patterns and insights on a specific topic, enabling researchers to gain a deeper understanding of complex phenomena. There are numerous examples of data-mining being coupled with data-analysis techniques to benefit domain-specific research (Abedin et al., 2010; Khan and McLeod, 2000; Jin et al., 2010; Haque et al., 2016; Jin et al., 2005; Ayoade et al., 2019). In the domain of conflict research, data mining techniques offer potential of analyzing vast

amounts of data related to conflict. This provides means for researchers to identify conflict patterns, escalation factors, root causes, and consequences, all of which can be used to help develop means to deescalate and handle present conflict. Additionally, data mining can provide insight into future conflict events, letting researchers forecast future political discord and proactively respond. Such an example is the Integrated Crisis Early Warning System, which utilizes automated even data to predict potential conflicts and conduct other types of political science research(Bagozzi et al., 2021; Beger et al., 2016; Brandt et al., 2022).

However, automated systems are not without limitations. Inaccuracy is common, and the context of a conflict may not be properly captured. This is in part due to automated systems relying heavily on pattern matching techniques and large dictionaries, which often yield poor results and are costly to maintain. Recent efforts by political scientist to remedy this issue include the use of traditional machine learning and deep learning techniques to better analyze political conflict and violence. (Hanna, 2017; Osorio et al., 2020; Beieler, 2016; Glavaš et al., 2017; Parolin et al., 2020). Unfortunately, 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.

A particular branch of NLP techniques, known as pre-trained language models, have been shown to be very effective on a variety of tasks that require a greater understanding of language as a whole. (Vaswani et al., 2017; Devlin et al., 2018b; Liu et al., 2019; Yang et al., 2019; Radford et al., 2019; **?**; Meta, 2023). Pre-training, in the context of NLP, is the act of providing a large corpus of unlabeled natural language to a model, allowing it to gain a broad and expansive understanding of language before it is ever trained on a specific task. The effectiveness of pre-trained language models can largely be attributed to recent

advancements in both computing power and data accessibility, enabling dense deep learning models to converge on mass amounts of text. Additionally, the construction of robust NLP benchmarks by researchers on a variety of tasks allows for an objective look at the performance of a particular model, allowing for further optimizations and developments in the field of NLP.

A common method of pre-training NLP models is via a generalized corpus, such as Wikipedia. (Zhu et al., 2015; Radford et al., 2019). However, recent research has shown the effectiveness of performing pre-training on domain-specific corpora (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). Specializing the pre-training corpus helps to bluster performance on domain-specific downstream tasks, such as natural language inference and question answering.

A pre-trained model by the name of ConfliBERT was shown to be effective in downstream tasks related to conflict(Hu et al., 2022). However, ConfliBERT has a limitation in that it can only be applied in English-contexts. In an effort to begin expanding ConfliBERT to a multilingual setting, ConfliBERT Spanish was developed. Spanish is one of the most spoken languages in the world(Statista, 2023). Additionally, numerous Spanish-speaking countries are experiencing increased political tension and conflict, such as Mexico, Peru, and Bolivia. (Carothers and Feldmann, 2021) Thus, Spanish was the most natural language to choose for multilingual expansion. ConfliBERT Spanish, similar to English variant, proved to be proficient in the domain of political conflict.

The downstream tasks that had been performed by ConfliBERT across all models and languages thus far were binary classification, named-entity recognition, and multi-label classification. In order to fully expand the capabilities of ConfliBERT, we decided that Extractive

Question Answering was the most natural additional for our models.

Extractive Question Answering (Extractive QA) is the NLP task of locating the answer of a question within a given context. Say you are given a Wikipedia page about the United States. This page would be considered the context. A question could be asking who the president of the U.S. is. Given these two as input, the output should be the starting and ending index of the context that answers the question. Extractive is only one type of QA task. Another common variant is generative QA, where the NLP model is tasked with generating its own answer instead of finding one in the given context. However, extractive QA was ultimately chosen over generative, as accuracy and factual basis is essential in the field of politics. An extractive QA model selects a section of text from a known and provided context; a generative QA model is forced to create its own answer, which is susceptible to hallucination and misinformation.

We propose both ConfliBERT English and Spanish, two pre-trained language models specifically tailored for Extractive Question Answering on conflict and political violence for their respective languages. 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 helped to curate a substantial corpus specifically tailored for English and Spanish language modeling within the domains of political violence, conflict, cooperation, and diplomacy.

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

- To exemplify the extractive question answering capabilities of ConfliBERT, we fine-tuned both our English and Spanish models on SQuAD v1.1 and NewsQA. SQuAD reflects performance in a generalized domain, while NewsQA shows conflict-domain extractive QA performance on a corpus consisting of news and politics.

- We performed thorough performance analysis on Extractive QA benchmarks to exemplify the significant increase of performance displayed by ConfliBERT English in comparison to models only trained on a general corpus.

- We heightened the accessibility of NLP in Spanish through our Spanish translation of NewsQA, our curated Spanish conflict corpus, and our ConfliBERT Spanish models. Being that Spanish is a low-resource language in the field of NLP, our work helps to provide tools and resources for further expansion of Spanish NLP research.

- We further expanded evaluation metrics of extractive QA models in the domain of conflict by creating ConfliQA, an extractive QA dataset consisting of 500 crowd-sourced QA pairs with relevant contexts, covering subjects such as ongoing conflict, war, policy, and human rights.

In Chapter 2 (Background), we will discuss necessary and relevant information in order to create a holistic view of the research we performed. In Chapter 3 (Approach), we will provide a full overview of our work, and how we approached it. In Chapter 4 (Experiment and Results), we will go into detail about how we performed our experiments, the results, and finally provide insight on the meaning behind the results. In Chapter 5, we will analyze our results thoroughly, suggest improvements and future work, and provide concluding remarks.

CHAPTER 2

BACKGROUND

2.1 NLP and Word Embedding

A fundamental aspect of most modern NLP models is word embedding. Put simply, word embedding is the act of converting words/symbols into a high-dimensional space. By doing so, we format natural language into an information-dense representation that is easily manipulable, allowing mathematics to be performed on normally non-numeric symbols. This is the building-block that enables machine learning to be performed on language, and is consequently why learning proper word-embeddings is such a critical task of language models.

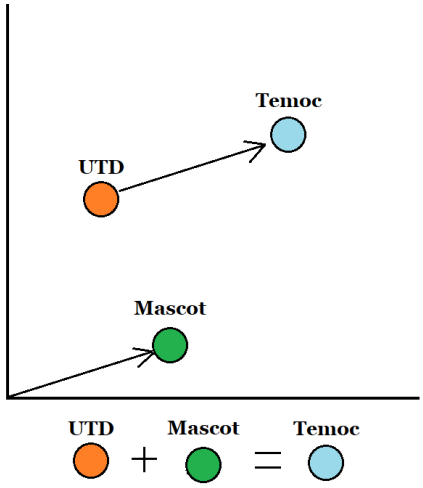

Figure 2.1. Example of Word Embedding

2.1.1 Non-contextual Embedding

Non-contextual embedding is a static embedding method, meaning that the embedding of a particular word does not change according to the context in which it exists. Non-contextual word embedding models, such as word2vec, (Mikolov et al., 2013) and Glove (Pennington et al., 2014), learn their embedding by mapping words to single vectors of a fixed length.

The resulting embedding represent each word as a dense vector of a fixed length. These vectors can then be used as input features for a wide range of NLP tasks, such as sentiment analysis (Hussein, 2018) and text classification (Kowsari et al., 2019).

Non-contextual benefits from being generally computationally efficient, requiring only a small amount of training data. However, the static nature of non-contextual embedding leads to difficulty capturing the meaning of words in different contexts. To overcome this limitation, contextual embedding was developed.

2.1.2 Contextual Embedding

Contextual embedding is a dynamic approach to embedding, varying the representation of words depending on the context in which they are used. Contextual embedding models form embeddings by mapping each word type to different vectors depending on the context of the sentence. The resulting embedding captures not only the meaning of a word, but also its relationship to other words within the context. This greatly benefits NLP tasks which require a more holistic perspective of the text being observed.

There has been a surge in contextual embedding models in recent years, particularly with the emergence of large pre-trained languages, such as BERT (Devlin et al., 2018b), RoBERTa (Liu et al., 2019), and GPT-3 (**?**). These models have demonstrated exceptional results on a wide range of NLP tasks, and have led to tremendous progression in NLP.

2.2 Pre-trained Language Models

In order to properly capture the complex relationships between words, a language model must be exposed to a large amount of text, the most popular method of which is pre-training. Pre-training entails initially providing a language model with a large corpus of unlabeled

text. Through pre-training, a language model is able to learn its own embedding, using the stream of text it receives as its input.

A tremendous advantage of using pre-training is that the process is entirely unsupervised. Language models observe the raw symbols within the text and produce their own embedding accordingly, without the need of labeling. This allows for mass amounts of text to be given during the pre-training process, as the bar to prepare text for the model is considerably low.

Early non-contextual embedded models, such as word2vec, (Mikolov et al., 2013) utilized pre-training to produce its embeddings. This posed a tremendous benefit, and enabled word2vec to perform well on downstream tasks such as sentiment analysis and text classification. However, as mentioned before, the limitations of using non-contextual embedding cause word2vec to struggle on NLP tasks that require a greater understanding of the contextual meaning and purpose of a word.

To address the limitations of non-contextual embedding, Embedding from Language Models (ELMo) was proposed (Peters et al., 2018). ELMo utilizes a bi-directional long-short term memory (LTSM) during its pre-training to extract contextual embeddings from the provided corpus. This proves to be a tremendous advantage, as it greatly outperforms previous non-contextual embedded models on NLP tasks that require a more holistic understanding of a context. However, despite the fact that its LTSM is bi-directional, it is a one way language model, leading it to struggle an understanding of the semantic information of a corpus. To address this, Google AI proposed the pre-trained Bidirectional Encoder Representations from Transformers (BERT), which implements bi-directional pre-training.

2.3 Transformers and Attention

Possibly the most significant advancement in NLP as of recent is the Transformer(Vaswani et al., 2017). The Transformer is a pre-trained NLP model that has two parts: the encoder and decoder. The Encoder takes in a sequence of tokens and uses multiple layers to generate encodings that represent the relationships between the tokenized inputs. In contrast, the Decoder performs the opposite, utilizing numerous decoding layers to convert an encoded sequence into an output. Perhaps the most significant aspect of the Transformer is its introduction of attention. This revolutionary mechanism allows the model to comprehend the relevance between tokens, greatly expanding the Transformer's ability to understand a complex context and the relationships between the words within it.

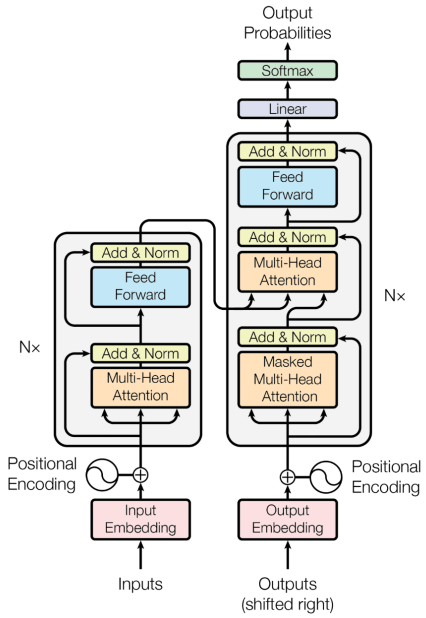

Figure 2.2. Diagram of Transformer
(Vaswani et al., 2017)

2.4 Bidirectional Encoder Representations from Transformers

All of the aforementioned NLP concepts culminate into Google's BERT(Devlin et al., 2018b). It is a pre-trained, Transformer-based language model capable of performing complex NLP tasks. The defining trait of BERT is its bidirectional approach to pre-training. When learning the embeddings of a word, BERT considers both the left and right context, resulting in rich contextual relationships between embeddings. There are two main approaches that BERT uses for its two-way joint training: Masked Language Model (MLM) and Next Sentence Prediction (NSP). In MLM, words within a context are randomly 'masked', and BERT is given the task of predicting what word the mask should be. Given that this forces the model to take in account purely the left and right context surrounding the mask, it helps to strengthen BERT's understanding of word relationships. NSP instead focuses on the relationships between sentences. NSP is given two sentences, and must determine if the second sentence is a continuation of the first. Where MLM strengthen understanding of words within a context, NSP heightens BERT's ability to understand the flow and bigger picture of larger spans of text.

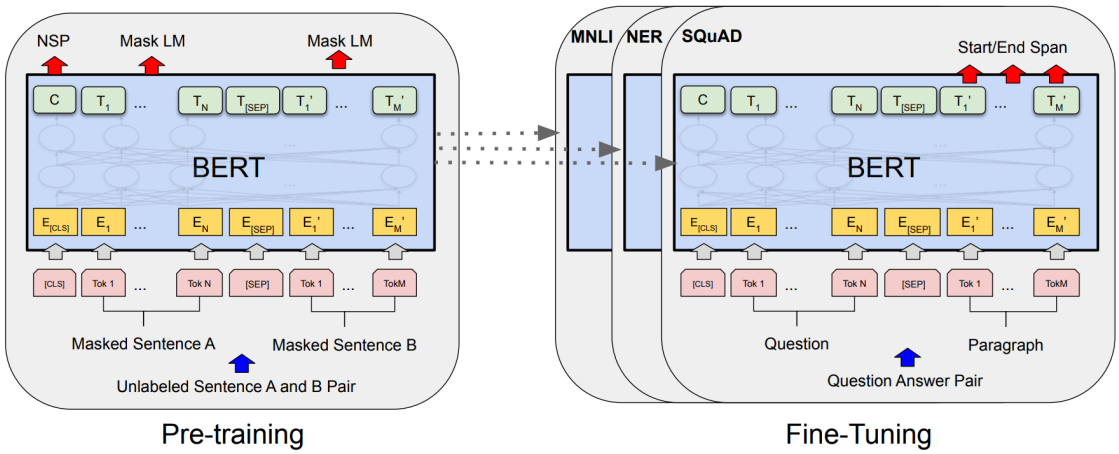

Figure 2.3. BERT Model Diagram
(Devlin et al., 2018b)

2.4.1 Variations of BERT

In response to the great success of the BERT, many variants of the PLM were developed. RoBERTa (Liu et al., 2019) (Robustly Optimized BERT approach) is a popular and highly successful variant of BERT, utilizing dynamic masking instead of static, and removing the use of NSP, both of which resulted in higher performance overall. ALBERTA (A Lite BERT) (Lan et al., 2019) utilizes a number of parameter-sharing techniques, helping to reduce the computational overhead of running the model while still maintaining equal performance to BERT. DistilBERT (Sanh et al., 2019) (Distilled version of BERT) takes on various design changed from the base BERT model in order to be faster to both train and deploy.

2.4.2 Domain-Specific BERT

Recent advancements in natural language processing (NLP) research have highlighted the substantial performance gains achieved by domain-specializing the corpora used during BERT pre-training for specific downstream tasks. BioBERT (Lee et al., 2020) is a BERT-based model that undergoes pre-training on a corpus of bio-medical literature, followed by fine-tuning on biology-specific tasks like bio-medical named entity extraction. Similarly, SciBERT (Beltagy et al., 2019) is pre-trained on a corpus of scientific publications and text, and fine-tuned on tasks such as scientific article classification and scientific question answering. LegalBERT (Chalkidis et al., 2020) utilizes pre-training on a diverse range of legal documents, and subsequent fine-tuning for legal document classification and named entity recognition of legal entities. Another notable model, ConfliBERT (Hu et al., 2022), is pre-trained on a carefully curated corpus comprising news articles, political texts, social media posts, and other conflict-related text. It is then fine-tuned for classification and named entity recognition of text associated with political conflicts. All of these specialized models have demonstrated significant performance improvements over the base BERT model, underscoring the value of domain-specific pre-training.

2.4.3 Extractive Question Answering

One of the most notoriously difficult downstream tasks for BERT is extractive question answering. To summarize, extractive question answering presents a language model with a context and a question. The context is a body of text about a specific topic, and the question pertains to the context. The job of the language model is to locate the answer to the question within the text. Popular extractive question-answering datasets include the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016), consisting of 100k+ question-answer pairs over various Wikipedia excerpts. The dataset is popularized for is significant size and generalized domain, allowing for ample data to fine-tune on.

CHAPTER 3

APPROACH

Following the success of ConfliBERT English (Hu et al., 2022) on conflict-related classification/ner tasks, efforts by UTD researchers were put into implementing the advancements mentioned in ConfliBERT's 'future work.' Two of these advancements were: the expansion of ConfliBERT's downstream tasks; the development of more, non-English ConfliBERT models. In this section, I will first describe our work on constructing our Conflict Corpus for the pre-training of ConfliBERT Spanish. Afterwords, I will explain the implementation of extractive question answering on both ConfliBERT English and ConfliBERT Spanish.

3.1 ConfliBERT Spanish

ConfliBERT Spanish, similar to ConfliBERT English, is a BERT-based LM that has been pre-trained on a large corpus of Spanish conflict-relevant text(Yang et al., 2023). After its pre-training, ConfliBERT Spanish was Fine-Tuned on Binary Classification, Multi-Label Classification, and Named Entity Recognition. Our contributions were in the data-collection of the model, where we scraped numerous domain-relevant websites from various Spanish-speaking countries.

Dataset	Domain	Task	Models	mBERT		BETO	
				cased	uncased	cased	uncased
Huffing tonpost	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**

Figure 3.1. ConfliBERT Spanish NER/BC/MLC Performance
(Yang et al., 2023)

3.1.1 Constructing Spanish Conflict Corpus

In order to create a BERT model capable of excelling at conflict analysis, we need to construct a corpus consisting of conflict-specific text. To fit this need, we developed our Spanish Conflict Corpus, which is comprised of 11.7GB of political conflict domain text. For our data acquisition, we predominantly scraped news websites, NGOs, and public Spanish datasets.

News Websites

In total, 123 unique news websites were scraped from across 18 Spanish-speaking countries. The motivation behind such a wide breadth of scraping was to prevent a regional bias in the information we scraped; such a characteristic could lead to ConfliBERT holding biases or perspective held by only a specific group of people. Additionally, we excluded any articles that were not classified as political texts, such as business and sports. Collectively, we

scraped 7.8GB of raw text.

Table 3.1 depicts information on the news websites that we scraped. The first column is the country in which the website was a part of. The second column depicts the name of the news website. The final column displays the size of text that was scraped.

Table 3.1. List of data acquisition sources-News websites

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 3.1 continued

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

Table 3.1 continued

Country	**News Website**	**Size**
Ecuador	Diario Los Andes	808 KB
	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 3.1 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

NGO Websites

A difficulty with scraping traditional news websites was filtering through content that was not conflict-related. Political scientists within our research team recommended scraping NGO websites, as they almost exclusively contain domain-specific text. Collectively, we scraped articles and text from 97 NGO websites across 8 Spanish speaking countries, totalling to roughly 1.1GB in size.

Table 3.2 depicts information on the NGOs that we scraped. They are in the same format as Table 3.1.

Table 3.2. 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 3.2 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 3.2 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

Public Spanish Datasets

Two public datasets of Spanish text were used in the curation of our corpora, namely MultiUN(Eisele and Chen, 2010) and DGT(Tiedemann, 2012). MultiUN is a multilingual corpus consisting of official UN documents translated into multiple languages. DGT is also a multilingual corpus, created by the European Union's Directorate-General for Translation. For both of these datasets, only Spanish text was used. The data we collected form these datasets totaled to 2.8GB.

Table 3.3. 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

Scraping Tools

In order to scrape these websites and datasets, we utilized a High Performance Computing (HPC) resource from the University of Arizona (UA). Python and Jupyter notebooks were utilized to automate the scraping of websites. Additionally, python libraries such as Newspaper, BeautifulSoup and Requests were used to extract relevant text from website endpoints. Regarding script development, our main contribution was an automated PDF scraper that used the PyPDF python library to extract text from pdfs, which enabled our team to access a much larger breadth of documents, such as papers or digital textbooks.

3.1.2 All ConfliBERT Spanish Models

Utilizing out conflict corpus, 4 ConfliBERT Spanish models were pre-trained, all of which being continually pre-trained from BERT-like models. The first two ConfliBERT Spanish models were continually pre-trained from Multilingual BERT Model(Devlin et al., 2018a).

Multilingual BERT is a BERT model that has been pre-trained on 104 different language, including Spanish. It has two variants: cased and uncased. The final two ConfliBERT Spanish models were continually pre-trained from BETO(Cañete et al., 2020), which is a BERT model pre-trained on entirely Spanish text. Similarly to Multilingual BERT, both a cased and uncased version were created.

Below are the names of the 4 ConfliBERT Spanish models:

- ConfliBERT-base-multi-cased
- ConfliBERT-base-multi-uncased
- ConfliBERT-BETO-cased
- ConfliBERT-BETO-uncased

3.2 Extractive Question Answering

Extractive question-answering was implemented on both ConfliBERT English and ConfliBERT Spanish, along with their base-model counterparts. Finding domain-specific extractive QA datasets to fine-tune on is challenging. Even more difficult is finding datasets that are parallel across multiple languages. Consequently, a large portion of this section will explore the translation of datasets from a starting language to a target language. Additionally, development of parallel conflict-specific evaluation will be explored.

3.2.1 BERT Extractive QA Fine-tuning Task

Extractive QA is one of many downstream tasks that BERT is able to perform, mostly due to its Transformer-Encoder architecture. To fine-tune extractive question answering,

BERT takes the Question and Context in as input, separating the two by the [SEP] token. After passing through all encoding layers, the output to the right of the [SEP] token will contain the tokens that BERT has determined are the 'answer' to the inputted question. Through numerous iterations and epochs, a BERT model is able to learn which tokens are of significance to answering the question.

3.2.2 Creation of Scripts

All ConfliBERT models and their respective base BERT counterparts are from the HuggingFace library. HuggingFace is a python library that provides extensive access to tools, datasets, and models related to the Transformer architecture. This made the process of Fine-Tuning on question-answering considerably easier, as HuggingFace provides courses and libraries specific to the downstream task. In the development of our fine-tuning, we utilized the HuggingFace library to create a python script that enables fine-tuning and evaluation of QA datasets. Additionally, it allows for parallelization across multiple GPUs, and lets the user batch multiple fine-tuning jobs by reading in multiple sets of arguments from json. To power our script, we utilized the National Center for Super-computing Applications at The University of Illinois, which provided me with 4 A100 GPUs, greatly increasing the speed and capabilities of our fine-tuning.

3.2.3 Translate Align Retrieve

As mentioned before, datasets for domain-specific QA are scarce; they're near non-existent when you consider a multilingual context. In order to combat this issue, a method known as Translate Align Retrieve (TAR)(Carrino et al., 2019) was developed. TAR is a Extractive Question-Answering translation method that has shown to be highly effective, and was responsible for translating SQuAD, a very popular QA dataset, into Spanish. First, it uses

machine translation to translate the Question, Answers, and Context into the target language. Next, it uses a cross-lingual word alignment model to map individual words in the original context to their most similar counterpart in the translated text. Finally, it uses the two models in conjunction to locate the starting and ending tokens of the answer within the Spanish context.

We used variation of this method to translate NewsQA, a CNN extractive QA dataset, into Spanish. Further details of our method will be described later on when discussing NewsQA as a whole.

3.2.4 Standardizing Dataset Format

In order to use the tools provided by HuggingFace, all datasets that weren't already formatted for HuggingFace, were. This meant using the python datasets library to configure the shape of a given dataset. The dataset would consist of two parts: Train and Validation. Train was used to fine-tune the model, and Validation was used to display the performance of the model after fine-tuning. Additionally, within both parts, there were 5 columns. The first column was id, which gave a unique string for each question-answer pair. Second was title, which categorized the context. Third was question, which contained the question to be asked as raw text. Fourth was context, which was the context as raw text. Fifth was answers, consisting of dictionaries with two keys: answer_start and text. Answer_start was a list of all the starting indexes of the answer within the context, while text was a list of raw text for each answer.

Each dataset's preparation was different, so individual python scripts were used to prepare datasets for BERT fine-tuning.

3.3 SQuAD v1.1

The Stanford Question Answering Dataset, or SQuAD(Rajpurkar et al., 2016), is an extractive question answering dataset developed by Stanford University. It consists of 100k+ question answer pairs regarding numerous Wikipedia excerpts over a generalized domain. The motivation behind using this dataset is that the performance of ConfliBERT can be compared to respective base models in order to observe if any performance on a general domain was lost from domain-specific pre-training. Additionally, this allows comparison to future conflict datasets, giving a baseline to determine any improvement gained when handling conflict-related QA tasks.

HuggingFace hosts a version of SQuAD that is easily accessible and compatible with HuggingFace QA model fine-tuning. Given this, no additionally data cleaning was required to prepare SQuAD for our BERT models, outside of preparing the fine-tuning script itself. HuggingFace also provides both versions of SQuAD: v1.1 and v2.0. v2 contains all rows from v1.1, but includes an additionally 50k rows that contain poor questions that cannot be answered. Noting this difference, SQuAD v1.1 was ultimately chosen over v2.0, as its performance on a generalized domain was deemed more relevant than the potential performance drop from being able to identify bad questions.

Although SQuAD was only released in English, there exist automatically translated versions in a variety of languages, the most well known being SQuAD_es(Carrino et al., 2019). SQuAD_es utilized the aforementioned TAR translation method to generate a high-quality automatic translation of SQuAD. As SQuAD and SQuAD_es are both parallel to eachother, we are able to perform equivalent performance evaluation across languages.

3.4 NewsQA

NewsQA is an English extractive question answering dataset developed by researchers at Microsoft(Trischler et al., 2017). It consists of over 100k crowd-sourced question answer pairs over CNN articles, many of which relate to political conflict and violence. However, due to copyright restrictions, the dataset is not immediately ready to be used, and must be constructed using various scripts provided by the creators. Additionally, once constructed, further data cleaning is required to remove bad question answer pairs, as well as to prepare it for HuggingFace libraries.

The NewsQA dataset was constructed using multiple waves of crowd-sourcing. The first wave was given CNN articles and was asked to write a question about them. The second wave was asked to provide answers to those questions, or mark a question as being bad. A third wave of crowd-sourcing was used to validate the quality of questions that did not have answer-overlap.

3.4.1 Construction

In conjunction with the NewsQA dataset, a link to a git repo is provided. CNN is unwilling to provide the explicit text of its articles in the dataset. As such, the dataset instead contains the article ids of each CNN article that is used. The scripts in the dataset map the article ids to the actual article text, generating the full dataset.

3.4.2 Data Cleaning

Despite the articles being properly aligned to the question-answer pairs, the dataset itself was still quite crude. The creators of the dataset decided to leave in all question-answer pairs, regardless of the quality denoted by the crowd-sourcers. Fortunately, they also included

information about the observations made by crowd-sources, such as the validated answers and poor questions/answers.

Before cleaning, there were roughly 120k question answer pairs. After removing all rows that were marked as having poor questions/answers, and only including validated rows, we were left with about 25k pairs. Although the quantity dropped by a significant factor, the quality of the remaining rows was considerably higher and much more relevant.

3.4.3 Formatting

To properly prepare the dataset for HuggingFace fine-tuning, we used a python Jupyter notebook and pandas. we eliminated any redundant rows, such as boolean rows indicating quality of questions. Additionally, we created new rows for id, title, question, context, and answer, which are the 5 required rows for HuggingFace. After formatting the dataset, we converted it to a datasetdict using the datasets python library. This is the same data format the SQuAD dataset is in; formatting them in an identical way helps to mitigate the risk of unexpected errors due to dataset discrepancies. Once NewsQA was a datasetdict, we stored it locally.

3.5 NewsQA Spanish

The only language that NewsQA is provided in is English, meaning that the dataset would have to be machine translated to Spanish. However, this is not a trivial task; alignment of words across languages is required to retrieve the new starting and ending answer indices for a given question context pair. To translate NewsQA into NewsQA Spanish, we utilized a variation of the TAR method mentioned earlier.

3.5.1 Translate

The original TAR method translates each context as a whole. However, the CNN articles in NewsQA are considerably larger than the contexts from SQuAD, and attempting to translate/align on them at this size could induce inaccuracies and noise. To combat this, we used the NLTK library to perform sentence tokenization, splitting the CNN articles up into individual sentences. From there, we used an English to Spanish NMT called opus-mt-en-es from HuggingFace, which provided parallel English to Spanish translations. We passed all sentences of a given CNN article at once to the NMT so that they could be translated in parallel. The result of this process was a list of translated sentences for each CNN Article. Additionally, each individual question was translated, although no sentence tokenization was used.

3.5.2 Align

The original TAR method uses fastalign(Dyer et al., 2013), a lightweight, unsupervised word aligner. However it was developed in 2013, and since then, higher quality word-alignment models have come into fruition. Instead of using fastalign, we chose to instead used SimAlign, a multilingual-bert based approach used to align source words to target words in a given sentence. Specifically, we utilized the ArgMax matching method, as it was shown to have performed the best in comparison to other methods available. Because CNN articles were translated per sentence, there were the same number of English sentences as there were Spanish sentences. So, instead of having to align the entire article, only the sentence(s) that contained the answer needed to be aligned. This greatly improved run-time and memory performance.

3.5.3 Retrieve

Sentences that contained the answer to a given question were aligned. Sometimes, the answer to a question could span multiple sentences. If this were the case, sentences were concatenated and aligned all-together. Once the alignments were made, all words in the original answer were replaced with their corresponding alignments in the translated text. Additionally, any 'gaps' in the answer were filled by highlighting the word with the lowest index to the word with the highest index. This left us with our translated answer. Finally, the starting index was found to the answer relative to the translated context, which was just the concatenation of all sentences for a given CNN story.

3.5.4 Formatting

The process of formatting the translated input was easier than the English NewsQA because the base of NewsQA Spanish was the cleaned and formatted NewsQA dataset. All that had to be done was to convert it back into the datasetdict format, which was trivial.

3.6 ConfliQA

ConfliQA is a crowd-sourced evaluation dataset for extractive QA models that we created for both English and Spanish. It is comprised of 500 QA pairs, parallel across both languages. It's context are from Wikipedia, covering global-wide conflict subjects, including war, ongoing conflicts, policy, human rights, terrorism, etc.

The motivation behind developing this dataset lies in the fact that there are no conflict QA datasets for Spanish that aren't entirely machine translated. With ConfliQA, we set out to create a dataset that could be answered and validated by native speakers of both English and Spanish. However, doing this entirely with crowd-sourcing would take extremely long.

Contrastly, using only machine learning and automation can lead to noise and incorrect information. With this in mind, we employed a methodology towards developing ConfliQA that used machine learning to do the busy work, leaving humans with more simple tasks, theoretically achieving the best of both worlds.

3.6.1 Methodology

The methodology for creating ConfliQA is as follows:

- 100 excerpts of conflict-domain Wikipedia articles were selected in English. If a parallel excerpt existed in Spanish, it was selected. Otherwise, Google Translate was used to translate the article into Spanish.

- For every excerpt, Google BARD (Bard, 2023)was provided the English text, and was prompted to create 5 questions, such that the questions could be answered within the context. The Questions were then auto-translated to Spanish using Google Translate.

- The 500 resulting question-context pairs were placed in a google document. The first wave of crowd-sourcers was tasked with highlighting the section of the context that contained the answer. If the answer could not be found, or the question/article were of poor quality, the crowd-sourcer was instead tasked with denoting the row accordingly.

- A second wave of crowd-sourcers was used to correct and repair poor questions and contexts, and then re-answer them.

- A third wave of crowd-sourcers was used to ensure that the questions, contexts, and answers remained parallel.

- A final wave of crowd-sourcers validated the quality of each question context answer pair.

This approach to developing ConfliQA was highly effective, and took only 3 days from start to finish. Theoretically, with more crowd-sourcing and a few tweaks to the methodology (which will be discussed later), this could be highly scaleable.

3.6.2 Construction

After the crowd-sourcing phase was complete, the dataset was exported locally and parsed into the same datasetdict format that SQuAD and NewsQA have been formatted into. It was then saved locally for future use.

CHAPTER 4

EXPERIMENT AND RESULTS

4.1 Fine-Tuning and Evaluation

Below are all models that were fine tuned and evaluated. All models used the same hyper-parameters and scripts. English models used English datasets, and Spanish models used Spanish datasets.

English Models

- BERT-base-cased
- BERT-base-uncased
- ConfliBERT-cont-cased
- ConfliBERT-cont-uncased
- ConfliBERT-scr-cased
- ConfliBERT-scr-uncased

Note that some models are denoted as 'cont,' while others are denoted as 'scr.' 'Cont' indicates that the model was continually trained from the base BERT-base model, while 'src' indicates that the model was build from scratch, using only the English Conflict Corpus. For Spanish ConfliBERT, all models were continually pre-trained from a base model, so 'scr' and 'cont' are not denoted.

Spanish Models

- BERT-base-multi-cased
- BERT-base-multi-uncased
- BERT-BETO-cased
- BERT-BETO-uncased
- ConfliBERT-base-multi-cased
- ConfliBERT-base-multi-uncased
- ConfliBERT-BETO-cased
- ConfliBERT-BETO-uncased

4.1.1 Metrics

When evaluating the performance of models, two metrics were used: exact match and F1 score. Exact match is the percent of answers that contain the exact beginning and ending tokens as the correct answer. F1 is more complex than EM, and provides insight on how well the model is capturing key-words in the answer. Below is the standard equation for F1.

$$F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

In this context, precision represents the percentage of tokens in the model's answer that can be found in the correct answer. Recall instead represents the percentage of tokens the model captured in its answer, out of all the tokens in the correct answer.

The motivation behind using these metrics is that they are largely standardized for extractive QA tasks. EM is able to capture the model's ability to adapt to the answering style of a QA dataset, while F1 is able to convey how much of the relevant text was captured in an answer.

4.1.2 Hyper-parameters

The following hyper-parameters were used for all experiments across all models:

- Seeds: 5
- Epochs: 4
- Batch Size: 64
- Learning Rate: 5e-5
- Max Answer Length: 30
- Max Content Length: 384
- Stride: 128

4.2 SQuAD v1.1

Below are the results from fine-tuning SQuAD v1.1 on all English and Spanish models. In all experiments, the portion of the dataset denoted as 'train' was used as the fine-tuning data, and the portion of the dataset denoted as 'validation' was used to benchmark performance.

4.2.1 ConfliBERT English

Table 4.1. English SQuAD Results

Model Name	F1 Score	Exact Match
ConfliBERT-cont-cased	**89.52**	**82.48**
ConfliBERT-cont-uncased	88.63	81.05
ConfliBERT-scr-cased	88.38	81.29
ConfliBERT-scr-uncased	88.79	81.68
BERT-base-cased	88.34	80.70
BERT-base-uncased	88.15	80.48

Analysis

Every ConfliBERT English model was able to outperform both base BERT models in both F1 and EM. The best performing was ConfliBERT-cont-cased. Given that SQuAD contains very little political information, it is possible that the additionally pre-training on-top of the base model's was enough to give a slight advantage to ConfliBERT on a generic-domain dataset.

4.2.2 ConfliBERT Spanish

Table 4.2. Spanish SQuAD Results

Model Name	F1 Score	Exact Match
ConfliBERT-base-multi-cased	76.60	62.57
ConfliBERT-base-multi-uncased	76.31	62.24
ConfliBERT-BETO-cased	**77.52**	**62.80**
ConfliBERT-BETO-uncased	74.12	59.72
BERT-base-multi-cased	76.76	62.65
BERT-base-multi-uncased	76.44	62.28
BERT-BETO-cased	77.10	62.66
BERT-BETO-uncased	74.33	60.01

Analysis

Similarly to English, ConfliBERT Spanish managed to outperform their respective base model counterparts, with both ConfliBERT-BETO-cased and BERT-BETO-cased performing best. This could suggest that the Spanish-only method of pre-training BETO models led to increased performance over the multilingual BERT models. As with English, this performance boost could possibly contributed to a larger pre-training corpus resulting in a slight improvement in comprehension of general domain text.

4.3 NewsQA

Below are the results from fine-tuning NewsQA English on all English models, and NewsQA Spanish on all Spanish models. In all experiments, the portion of the dataset denoted as 'train' was used as the fine-tuning data, and the portion of the dataset denoted as 'validation' was used to benchmark performance.

4.3.1 ConfliBERT English

Table 4.3. English NewsQA Results

Model Name	F1 Score	Exact Match
ConfliBERT-cont-cased	71.20	49.03
ConfliBERT-cont-uncased	71.56	49.78
ConfliBERT-scr-cased	**72.90**	**50.65**
ConfliBERT-scr-uncased	71.22	49.05
BERT-base-cased	68.91	45.67
BERT-base-uncased	68.90	46.39

Analysis

As expected, when ConfliBERT English is presented with a conflict-domain task, it outperforms base BERT by a significant margin. The performance improvement is also more significant than seen on SQuAD, further emphasizing this point.

4.3.2 ConfliBERT Spanish

Table 4.4. Spanish NewsQA Results

Model Name	F1 Score	Exact Match
ConfliBERT-base-multi-cased	64.09	34.51
ConfliBERT-base-multi-uncased	63.66	33.71
ConfliBERT-BETO-cased	66.77	**36.61**
ConfliBERT-BETO-uncased	62.68	32.75
BERT-base-multi-cased	63.54	33.32
BERT-base-multi-uncased	63.49	33.73
BERT-BETO-cased	**66.82**	36.38
BERT-BETO-uncased	62.54	33.78

Analysis

In great contrast to ConfliBERT English, ConfliBERT Spanish actually remains incredibly close to the base BERT models. ConfliBERT-BETO-cased manages to outperform on EM, but BERT-BETO-cased manages to obtain a better F1 score. This could indicate that the conflict corpus that was curated for ConfliBERT was not large enough to have a significant impact on performance. The English Conflict Corpus was 34GB, while the Spanish Conflict Corpus was only 11.7GB. This is in part due to the difficulty of mining political-specific text. Regardless, it is clear that ConfliBERT Spanish is in need of improvements in order to be relevant for extractive QA in the domain of political conflict. Further analysis will be done in Chapter 5, where changes that can be made to ConfliBERT to increase scores on extractive QA downstream tasks will be discussed.

4.4 ConfliQA

Below are the results from evaluating all fine-tuned models on ConfliQA. In all experiments, the portion of the dataset denoted as 'evaluation' was used to benchmark performance.

4.4.1 ConfliQA English Fine-Tuned on SQuAD v1.1

Table 4.5. Results of ConfliQA English Fine-Tuned on SQuAD v1.1

Model Name	F1 Score	Exact Match
ConfliBERT-cont-cased	**70.52**	46.00
ConfliBERT-cont-uncased	69.80	45.80
ConfliBERT-scr-cased	70.35	**47.20**
ConfliBERT-scr-uncased	68.91	44.60
BERT-base-cased	69.08	43.60
BERT-base-uncased	69.98	46.00

Analysis

Despite the significant performance observed when ConfliBERT was given NewsQA, a conflict-domain extractive QA dataset, the performance gap between ConfliBERT English and the base bert models on ConfliQA is fairly minor. This is most likely indicating the importance of not only pre-training on a conflict-domain corpus, but also fine-tuning on one as well.

4.4.2 ConfliQA Spanish Fine-Tuned on SQuAD es v1.1

Table 4.6. Results of ConfliQA Spanish Fine-Tuned on SQuAD es v1.1

Model Name	F1 Score	Exact Match
ConfliBERT-base-multi-cased	63.35	28.51
ConfliBERT-base-multi-uncased	63.98	31.18
ConfliBERT-BETO-cased	64.29	30.58
ConfliBERT-BETO-uncased	61.88	30.78
BERT-base-multi-cased	65.02	30.78
BERT-base-multi-uncased	64.96	30.18
BERT-BETO-cased	**66.95**	**33.80**
BERT-BETO-uncased	62.75	29.97

Analysis

ConfliBERT Spanish fine-tuned on SQuAD underperforms on ConfliQA, which is most likely the joint-issue of not having enough conflict-related pre-training, and not being fine-tuned on a conflict-related extractive QA dataset. However, ConfliBERT-BETO-cased has consistently been performing best for ConfliBERT and BERT base models across all tasks. This could indicate that future work on Spanish ConfliBERT may benefit from using an all-Spanish pre-training corpus, such as with BETO.

4.4.3 ConfliQA English Fine-Tuned on NewsQA

Table 4.7. Results of ConfliQA English Fine-Tuned on NewsQA

Model Name	F1 Score	Exact Match
ConfliBERT-cont-cased	59.81	32.20
ConfliBERT-cont-uncased	59.09	31.00
ConfliBERT-scr-cased	59.69	30.20
ConfliBERT-scr-uncased	**60.84**	**32.40**
BERT-base-cased	57.72	28.60
BERT-base-uncased	57.75	30.00

Analysis

By fine-tuning ConfliBERT English on a conflict-domain extractive QA dataset, we see a much greater performance gap between the base models, with ConfliBERT winning by a sizable margin. However, we also notice that the F1 and EM scores are lower than when we fine-tuned them on SQuAD for both ConfliBERT and base BERT models. This is likely due to the fact that SQuAD is formatted much more similarly to ConfliQA than NewsQA. SQuAD and ConfliQA are both use Wikipedia excerpts as their contexts, while NewsQA uses entire CNN articles, and often contains very large spans of text as its answers. This indicates that the style of questions/contexts you are planning to use your model on are an important factor when deciding what fine-tuning dataset to use.

4.4.4 ConfliQA Spanish Fine-Tuned on NewsQA Spanish

Table 4.8. Results of ConfliQA English Fine-Tuned on NewsQA

Model Name	F1 Score	Exact Match
ConfliBERT-base-multi-cased	55.38	**12.87**
ConfliBERT-base-multi-uncased	51.13	10.06
ConfliBERT-BETO-cased	56.57	12.07
ConfliBERT-BETO-uncased	49.87	8.85
BERT-base-multi-cased	53.99	12.07
BERT-base-multi-uncased	55.21	10.66
BERT-BETO-cased	**57.53**	11.87
BERT-BETO-uncased	51.08	10.26

Analysis

ConfliBERT Spanish fine-tuned on NewsQA shows a similar behavior to its English counterpart in that that performance gap between ConfliBERT and base models improves as opposed to with SQuAD, but the overall scores lower. Unfortunately, this improvement is not enough to prove its effectiveness on conflict-domain QA tasks, as only 1 of the 4 ConfliBERT models outperforms its respective base model.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Analysis of Results

On all tests, English ConfliBERT has outperformed its base-model counterpart, exemplifying the effectiveness of a large pre-training corpus of political conflict text on conflict-domain extractive QA, as well as the competence of the model. However, the performance of Spanish ConfliBERT was considerably more inconsistent.

This is most likely due to the size of the Spanish Conflict Corpus used for pre-training in comparison to the English Conflict Corpus; the English corpus was nearly 3 times the size at 34GB, while Spanish was at 11.7GB. The smaller size was predominately due to the amount of time and effort it took to find relevant text for the corpus.

Additionally, all of the Spanish fine-tuning datasets that were used, SQuAD Spanish and NewsQA English, were all translated. The use of NMT translation could be introducing noise or inaccurate information, and the alignments for answers could contain inaccuracies.

5.2 Improving ConfliBERT Spanish for Extractive QA

There are a number of improvements that can be made to increase performance. They largely fall into three categories: increasing corpus size, developing our own fine-tunable dataset, and translating more datasets with improved methodologies.

5.2.1 Increasing Corpus with more Data Mining

The most simple and direct solution is to increase the conflict corpus used for pre-training. A subset of political websites that we did not have time to mine were government websites;

often times, government websites deal almost entirely in politics/human rights, which would be incredibly relevant for conflict.

Additionally, improvements to our automatic crawling would greatly decrease the amount of labor expended per website. As of now, we have to tailor our scripts per website. An idea to improve this could be to use our existing political text classification task with ConfliBERT Spanish, which has shown great results. A heuristic-first search algorithm could be devised to search websites through hrefs, and prioritize those that contain relevant political text.

5.2.2 Developing a Fine-Tunable Dataset

A significant drawback with our approach to fine-tuning ConfliBERT Spanish on conflict extractive QA was the lack of non-machine translated datasets. A tremendous contribution to conflict research would be the development of our own, crowd-sourced conflict dataset, large enough to be fine-tuned.

However, as discussed in previously, the amount of effort and money that goes into crowd-sourcing can make it impractical. In order to develop our own fine-tunable dataset, we would need to employ both NLP and human evaluation in tandem.

We propose the following:

- Find numerous relevant contexts in the target language. Ensure that they are not translated. This process can be automated largely by scraping paragraphs from known political sources, such as news websites, Wikipedia, etc.

- Use a generative model, such as GPT, BARD, etc. to create relevant, extractive questions. Create 3-10 questions per context depending on the length and content within.

- Use a fine-tuned model, such as multilingual BERT, to extract answers from the context question pairs. Take the top 5 answers for each question.

- Provide crowd-sourcers the task of selecting which answers are correct. If no answers are correct, mark as such. If no answers are selected, this means either that the model failed to find the answer, or the question is poor. Have the crowd sourcers denote which it is.

- For each bad question, have a second wave of crowd sourcers create a better question that doesn't match any other questions used on the given context.

- For each bad answer, have the second wave of crowd-sourcers highlight correct answers.

Using question generation and EQA fine-tuned BERT to answer the questions provides a significantly easier task for crowd-sourcers, enabling them to work much faster. The subset of bad questions/answers will be significantly smaller than the entire dataset, reducing the number of pure-human answers needed, while still ensuring quality.

Alternatively, you could instead replace the second wave of crowd-sourcers with more NLP/generation, but this could introduce a bias/noise.

5.2.3 Translating More Datasets

If it is done well, translating QA datasets is an extremely effective way to expand low-resource languages. Besides TAR, another method used by the datasets MLQA (Lewis et al., 2019)

is known as quoting.

Essentially, before translation, you encase the answer in a pair of quotes, which should theoretically transfer over across the translation. However, there arise a few issues. First, the quotes do not always survive the translation. Secondly, news/conflict related text often contains quotations, so deciding which set of quotes would be an additional challenge. Other symbols could be used, such as brackets or parenthesis, but they may not perform as well as quotes.

An alternative method that we came up with involves fine-tuning BERT on an extractive downstream task. Essentially, you give multilingual BERT a your context in the translated language, and an answer in the language you started with. The goal is for BERT to take the starting answer, and find its translated equivalent in the translated text. An effective way creating a dataset for this task would be to take a parallel extractive QA dataset, and use the context and answer from the target language dataset, and only the answer from the starting language dataset. We unfortunately did not have enough time to employ this technique, but we plan on experimenting with it more in the future.

5.3 Future Work

Although I am graduating from UTD this semester, I plan on continuing my research while I am away. I also plan on returning for a Ph.D in the coming years. NLP is an exciting field, and there is certainly more work to be done.

- Implementation of the improvements to ConfliBERT Spanish should be a priority. Additionally, a ConfliBERT Arabic model is currently being developed. It should also be further expanded to extractive QA.

- Expanding ConfliBERT to more tasks, such as generative QA, summarizing, and more, would provide conflict researchers with even more tools at their disposal.

- The Conflict corpora that we have created are not exclusive to only BERT; exploration of other models and their use cases is essential to staying relevant in NLP.

5.4 Conclusion and Contribution

Through our research into conflict-domain extractive QA, we have been able to greatly expand the capabilities of ConfliBERT English, proving its effectiveness on conflict-related QA, providing conflict researchers with more powerful NLP tools. Additionally, we helped to develop ConfliBERT Spanish, identified ways in which we can improve it, and showed further insight on how to approach other low-resource languages. We further expanded fine-tunable datasets in the domain of Spanish political science QA by cleaning and translating NewsQA to Spanish with a modern, refined approach. The scripts used to translate NewsQA can be used for any target language so long as NMTs and alignment methods exist for those

languages. Finally, we developed ConfliQA, which provides a reliable and clean benchmark for fine-tuned extractive QA LMs, and led to further discovery of new ways to effectively create QA datasets.

REFERENCES

Abedin, M., V. Ng, and L. Khan (2010, August 26). Cause identification from aviation safety incident reports via weakly supervised semantic lexicon construction. *Journal of Artificial Intelligence Research 38*, 569–631.

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*.

Ayoade, G., A. El-Ghamry, V. Karande, L. Khan, M. Alrahmawy, and M. Z. Rashad (2019, August 1). Secure data processing for iot middleware systems. *The Journal of Supercomputing 75*, 4684–4709.

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.

Bard, G. A. (2023). Citing bard outputs.

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.

Carothers, T. and A. Feldmann (2021). Divisive politics and democratic dangers in latin america.

Carrino, C. P., M. R. Costa-jussà, and J. A. R. Fonollosa (2019). Automatic spanish translation of the squad dataset for multilingual question answering.

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*.

Devlin, J., M. Chang, K. Lee, and K. Toutanova (2018a). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805*.

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

Dyer, C., V. Chahuneau, and N. Smith (2013). A simple, fast, and effective reparameterization of ibm model 2. In *NAACL Proc 2013*.

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).

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.

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

Haque, A., L. Khan, and M. Baron (2016, February 21). Sand: Semi-supervised adaptive novel class detection and classification over data stream. *Proceedings of the AAAI Conference on Artificial Intelligence 30*.

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.

Jin, Y., L. Khan, and B. Prabhakaran (2010, March). Knowledge based image annotation refinement. *Journal of Signal Processing Systems 58*, 387–406.

Jin, Y., L. Khan, L. Wang, and M. Awad (2005, November 6). Image annotations by combining multiple evidence & wordnet. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pp. 706–715.

Khan, L. and D. McLeod (2000, August 20). Effective retrieval of audio information from annotated text using ontologies. In *Proceedings of the First International Conference on Multimedia Data Mining*, pp. 37–45.

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*.

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.

Lewis, P., B. Oğuz, R. Rinott, S. Riedel, and H. Schwenk (2019). Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

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.

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 (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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.

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., 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.

Pennington, J., R. Socher, and C. D. Manning (2014). 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*.

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

Rajpurkar, P., J. Zhang, K. Lopyrev, and P. Liang (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints*, arXiv:1606.05250.

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.

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.

Statista (2023). The most spoken languages worldwide in 2023(by speakers in millions).

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

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

Trischler, A., T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman (2017). Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pp. 191–200.

United Nations (2020). A new era of conflict and violence. *United Nations 75*.

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

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, W., S. Alsarra, L. Abdeljaber, N. Zawad, Z. Delaram, P. Whitehead, J. Osorio, L. Khan, P. Brandt, and V. D'Orazio (2023). Conflibert-spanish: A pre-trained spanish language model for political conflict and violence. In *5th IEEE Conference on Machine Learning and Natural Language Processing: Models, Systems, Data and Applications*.

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*.

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

Parker Whitehead began attending The University of Texas at Dallas in 2019. He was accepted into the Fast-Track program in 2021, allowing the pursuit of a bachelor's and master's at the same time. He graduated Summa Cum Laude from UTD in May of 2022 with a Bachelor of Computer Science. Currently, his research interest is in Natural Language Processing, and he plans on graduating with a Master of Computer Science from UTD in August of 2023. After graduating, he plans on pursuing a PhD in NLP after a few years of work in the industry.

CURRICULUM VITAE

Parker Whitehead

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: pmw180000@utdallas.edu

Education:

BS, Computer Science, The University of Texas at Dallas, 2022

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

Employment History:

Graduate Software Engineer & ML Intern, MKS Instruments, Summer 2022
Software and Machine Learning Engineer, MKS Instruments, August 2022 - Present