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

Yibo Hu[†], **MohammadSaleh Hosseini**[†], **Erick Skorupa Parolin**[†],
Javier Osorio[§], **Latifur Khan**[†], **Patrick T. Brandt**[†], **Vito J. D'Orazio**[†]
[†]The University of Texas at Dallas, [§]The University of Arizona
{yibo.hu,seyyedmohammadsaleh.hosseini,erick.skorupaparolin,
lkhan,pbrandt,dorazio}@utdallas.edu,josorio1@email.arizona.edu

Abstract

Analyzing conflicts and political violence around the world is a persistent challenge in the political science and policy communities due in large part to the vast volumes of specialized text needed to monitor conflict and violence on a global scale. To help advance research in political science, we introduce ConfliBERT, a domain-specific pre-trained language model for conflict and political violence. We first gather a large domain-specific text corpus for language modeling from various sources. We then build ConfliBERT using two approaches: pre-training from scratch and continual pre-training. To evaluate ConfliBERT, we collect 12 datasets and implement 18 tasks to assess the models' practical application in conflict research. Finally, we evaluate several versions of ConfliBERT in multiple experiments. Results consistently show that ConfliBERT outperforms BERT when analyzing political violence and conflict. Our code is publicly available.[1]

1 Introduction

The study of political violence is a central concern of conflict scholars and security analysts in the academic and policy communities. For decades, scholars and governments have devoted incalculable resources to monitoring, understanding, and predicting the dynamics of social unrest, political violence, and armed conflict worldwide. Conflict research is a sub-field of political science that analyzes a broad scope of interactions between government agents, their challengers, and the civilian population, including material and verbal conflict and cooperation. Conflict research covers protest, riots, repression, insurgency, civil war, terrorism, human rights, genocide, criminal violence, forced displacement, conventional and unconventional warfare, nuclear deterrence, peacekeeping, diplomatic disputes and cooperation, among others.

[1] https://github.com/eventdata/ConfliBERT

Traditionally, researchers used manual coding to track conflict processes worldwide (Raleigh et al., 2010). Unfortunately, the high costs and slow pace of domain experts conducting these tasks make it extraordinarily difficult and costly to monitor highly complex and rapidly changing conflicts in an ever-growing volume of information available on a global scale. Furthermore, these efforts tend to focus on quantifying particular types of conflict events between specific kinds of actors (Sundberg and Melander, 2013).

Initial efforts to address these challenges motivated political scientists to develop automated systems to classify or extract structured event data from news articles (Bond et al., 2003; Boschee et al., 2016; O'Brien, 2010; Osorio and Reyes, 2017; Schrodt, 2006, 2009; Alliance, 2015; Norris et al., 2017; Lu and Roy, 2017; Ward et al., 2013). These systems capture a broader range of event types, including different conflict and cooperation events, between a larger number of political actors. They can also extract volumes of data that are orders of magnitude greater than manual coding efforts. Automated event data such as the Integrated Crisis Early Warning System have been used for conflict forecasting and other kinds of research in political science (Bagozzi et al., 2021; Beger et al., 2016; Brandt et al., 2022).

However, these existing systems rely on dated pattern matching techniques and large dictionaries, which often yield low-accuracy results and are too costly to maintain. Recent efforts by political scientists employ traditional machine learning (Hanna, 2017; Osorio et al., 2020) and deep learning (Beieler, 2016; Radford, 2020b; Glavaš et al., 2017; Skorupa Parolin et al., 2020) to analyze political conflict and violence. Standard supervised learning requires labeled data, which are expensive to obtain due to the expertise required for quality annotation. This led conflict scholars to seek alternative solutions based on the latest developments

in natural language processing (NLP).

Recent progress in NLP has been driven by pre-trained transformer language models (Vaswani et al., 2017; Radford et al., 2019; Devlin et al., 2018; Yang et al., 2019). Self-supervision using large-scale unlabeled text can significantly alleviate the annotation bottleneck using transfer learning. The training parallelization of transformers also improves their efficiency on large datasets. As a result, the use of powerful computational devices and the advantage of transformer structures make large-scale language models' pre-training possible. Furthermore, the introduction of extensive benchmarks (Wang et al., 2018, 2019; Rajpurkar et al., 2018; Lai et al., 2017) validates the significant improvement of pre-trained language models on various downstream tasks.

While many language models are built on general domain corpora, such as Wikipedia, BookCorpus (Zhu et al., 2015), and WebText (Radford et al., 2019), recent works show that pre-training on domain-specific corpora can boost downstream performance on those domains (Lee et al., 2019; Gururangan et al., 2020). Domain-specific work in bio-medicine focuses not only on developing pre-trained models (Lee et al., 2019; Beltagy et al., 2019; Alsentzer et al., 2019; Lewis et al., 2020; Gu et al., 2021) but also on proposing domain-relevant evaluation benchmarks (Peng et al., 2019; Gu et al., 2021). Pre-training models also have advanced research in other domains such as academic papers (Beltagy et al., 2019) and legal studies (Chalkidis et al., 2020). Despite some efforts to apply transformers-based approaches in political science (Büyüköz et al., 2020; Olsson et al., 2020; Örs et al., 2020; Radford, 2020a; Halterman and Radford, 2021; Hürriyetoğlu et al., 2021; Parolin et al., 2021a, 2022), we are unaware of any studies that develop and evaluate domain-specific pre-trained language models for political science or conflict research.

By combining the expertise of conflict scholars and computer scientists, we developed ConfliBERT, a pre-trained language model designed for conflict and political violence. ConfliBERT improves downstream tasks for conflict research while significantly alleviating the annotation bottleneck. We expect it to support a broad community of academic and policy researchers, enabling the analysis of conflict processes using a domain-specific NLP tool that yields accurate and valid results at minimum operational cost. Our paper provides the following key contributions: (1) We curate a large domain-specific corpus for language modeling in the domain of political violence, conflict, cooperation, and diplomacy. (2) Based on our domain-specific corpora, we devise a pre-trained language model, ConfliBERT, and make it available to the general public, which directly benefits the political science and policy communities. (3) To evaluate our model in practical applications, we collect 12 datasets and conduct 18 tasks relevant to conflict research. We are the first to carry out such a comprehensive evaluation of language models for conflict studies. (4) We evaluate different versions of ConfliBERT and show it outperforms models trained on generic domains. We also perform in-depth analyses of different tasks to investigate the factors affecting the performance.

2 Preliminaries

Recent pre-trained transformer language models, such as the Bidirectional Encoder Representation from Transformers (BERT) (Devlin et al., 2018), follow a two-steps framework: (1) pre-train on a large unlabeled corpus; and (2) fine-tune on task-specific labeled data. These models learn semantics during the pre-training step and require smaller labeled data to significantly improve their performance on downstream tasks. The fine-tuning step requires minor network modifications to create state-of-the-art models for various tasks.

Technically, BERT uses the multi-layer, multi-head self-attention mechanism, which provides substantial advantages for language modeling, such as allowing parallel GPU computation and capturing long-range dependencies. This allows to efficiently pre-train large language models on large corpora using powerful devices.

Another key element of BERT-like models is the design of self-supervision tasks. Self-supervision refers to generating labels for unlabelled data and using them to train a model in a supervised manner. BERT uses two self-supervision tasks during pre-training. On the one hand, masked language model (MLM) is a fill-in-the-blank task based on randomly masking a token and then using the surrounding words to predict the word hidden behind the mask. On the other hand, next sentence prediction (NSP) determines whether one sentence follows another one in the same document.

Recent works also propose variants of self-

supervision tasks. For example, Joshi et al. (2020) mask out contiguous sequences of tokens to improve span representations. Clark et al. (2020) use replaced token detection, where the model distinguishes real input tokens from plausible but synthetically generated replacements. However, Liu et al. (2019) prove that MLM is competitive with other recently proposed training objectives with more data and improved training strategies.

Finally, most BERT-like models focus on a generic domain, such as Wikipedia, BookCorpus (Zhu et al., 2015) and WebText (Radford et al., 2019). However, BERT without domain adaptation tends to underperform in target domains with distinct characteristics such as specialized vocabulary, language style, and specific semantics. Our domain includes political violence, armed conflict, international cooperation, and diplomacy—all of which have these characteristics. This performance gap is the primary motivation for developing domain-specific language models.

Specifically, the language of political actors involves strategic and complex semantics. Policy positions that show support for one actor while also threatening another are sometimes embedded in simple statements. For example, "NATO will not tolerate this aggression" mixes a negation, a conditional, and the action of potential interest. Signals are highly context-dependent, adapted for a target audience, and vary in strength depending on the specific actor sending the signal (McManus, 2017; Blankenship, 2020). Compared to a generic language model, we expect ours to incorporate important contextual information and learn more about the strategic ways that political actors convey information. This context and *the related political biases in it* are exactly a need that ConfliBERT aims to fulfill. Political conflict and violence text gains from this domain knowledge: one political actor's definition of "rebels" is another's "freedom fighters".

Table 1 summarizes various recent domain-specific BERT models, including our model, ConfliBERT. These models mainly differ in their corpora and pre-training strategies, including: (1) continuing pre-training (**Cont**); and (2) pre-training from scratch (**SCR**). In the next section, we elaborate on the strategies and our method of developing ConfliBERT in the domain of political conflict and violence.

Model	Method	Corpora and Text Size
BERT	-	Wiki+Books, 3.3B words/16 GB
BioBERT	Cont	PubMed, 4.5B words
SciBERT	SCR	BIO+CS papers , 3.2B words
BlueBERT	Cont	PubMed+MIMIC, 4.5B words
PubBERT	SCR	PubMed, 3.1B words/21 GB
LegalBERT	Both	legislation, court cases, 12 GB
ConfliBERT	Both	organization/government reports, news, 7B words/34 GB

Table 1: Summary of selected BERT models in general and specific domains.

3 Approach

As described in Section 2, MLM-based BERT achieves competitive performance among other transformer models with different self-supervision tasks. Besides, BERT has been validated in various domains (Lee et al., 2019; Beltagy et al., 2019; Peng et al., 2019; Chalkidis et al., 2020; Gu et al., 2021) shown in Table 1. Therefore, we develop our domain-specific model based on BERT. The key components of developing and validating our model, ConfliBERT, include pre-training strategies, corpora, and evaluation tasks.

3.1 Domain-specific Pretraining

We explore both strategies (SCR and Cont) of adapting BERT to the political conflict and violence domain. A Cont model initializes with BERT's checkpoint and vocabulary, and trains for additional steps on a domain-specific corpus. Since BERT has already been pre-trained about one million steps on the generic domain, Cont usually requires fewer steps than training a new model from scratch. For example, Lee et al. (2019) report that continual pre-training of BERT on a biomedical dataset for 470K steps yields comparable performance to pre-training for one million steps.

On the other hand, when pre-training BERT from scratch (SCR) on the domain-specific corpora, we generate a new vocabulary from the target domain instead of using the original BERT's vocabulary. Various papers (Beltagy et al., 2019; Gu et al., 2021) argue that SCR generates substantial gains over Cont for domains with sizeable unlabeled text.

We refer to the original BERT vocabulary as BaseVocab and our domain vocabulary as ConfliVocab. We generated both cased and uncased versions of ConfliVocab on our training corpus using the Wordpiece algorithm (Wu et al., 2016). We set the ConfliVocab size to 30,000 words to

Words	BERT	ConfliBERT
Daesh	Da-esh	Daesh
extremists	ex-tre-mist-s	extremists
FARC	FA-RC	FARC
IED	I-ED	IED
indiscriminately	in-dis-c-rim-inate-ly	indiscriminately
manhunt	man-hun-t	manhunt
mutilation	m-uti-lation	mutilation
paramilitaries	para-mi-lit-aries	paramilitaries
perpetrator	per-pet-rator	perpetrator
punitive	pu-ni-tive	punitive
racketeering	rack-ete-ering	racketeering
separatists	se-par-ati-sts	separatists
subversive	sub-vers-ive	subversive
undemocratic	und-em-oc-ratic	undemocratic
xenophobic	x-eno-phobic	xenophobic

Table 2: Examples of common terms in conflict domain.

match that of BaseVocab. The resulting token overlap between BaseVocab and ConfliVocab is 58.3%, which indicates a considerable difference (41.7%) in high-frequency words between the general and conflict-specific corpora.

In particular, we find a substantial advantage of using ConfliVocab during the tokenization. Table 2 shows examples of conflict-related terms that exclusively appear in ConfliVocab. For example, the term "separatists" is not included in BaseVocab, and BERT erroneously splits it into four sub-words ["se", "##par", "##ati", "##sts"]. This fragmentation may hinder learning in downstream tasks. We will validate the advantage of ConfliVocab in the downstream tasks in our experiments section.

3.2 Corpora

The first step to develop ConfliBERT is to build a domain-specific corpus for pre-training. As illustrated in Table 1, there exist large-scale publicly available biomedical datasets, such as PubMed and MIMIC (Johnson et al., 2016). SciBERT (Beltagy et al., 2019) is built from a large corpus of academic papers (Ammar et al., 2018; Lo et al., 2020). History Lab[2] provides many government documents, but lacks the breadth we need for the politics and conflict domain (Connelly et al., 2021). Thus we curated a domain corpus that consists of 33.7 GB of clean, plain text in the BERT required format. We bin the sources into five categories below and provide more details in Appendix.

Expert Domain Corpora (EDC). We curated 2,293 MB of plain text from multiple professional sources relevant to conflict and diplomacy. The sources include United Nations' websites and databases, international humanitarian non-governmental organizations, think tanks, and government sources such as the Foreign Relations of the United States. These are examples of *objective records of government and diplomatic activity from non-partisan observers.*

Mainstream Media Collection (MMC). We crawled 35 worldwide news agencies reporting in English and with coverage from 1966 to 2021. We pre-processed and filtered 20 GB of stories using metadata such as document tags for War and Politics. These cover a period during and after the Cold War with global coverage that focuses on primarily state-based conflict.

Gigaword. This corpus provides a distinct coverage of seven international English newswires from 1994 to 2010 (Parker et al., 2011). We removed the overlapping stories (which also existed in MMC) and filtered an 8,818 MB domain-specific subset.

Phoenix Real-Time (PRT). PRT is a developing event dataset crawled from more than 400 news agencies worldwide from October 2017 (Salam et al., 2018). It contains many news agencies in areas other than Europe and the U.S., thus improving the scope of our coverage. We removed the duplicated news agencies (which also existed in MMC and Gigaword) and filtered a 2,425 MB relevant subset. This allows the capture of post-Cold War actors, the Global War on Terrorism Service Medal (GWOT), and more recent events.

Wikipedia. Wikipedia has a different language style for describing political events and can enrich the diversity of our corpus. Based on its category labels, we curated 2,845 MB of relevant articles from an 18 GB size of the Wikipedia dump released on March 20, 2021.

3.3 Evaluation Tasks

The introduction of comprehensive benchmarks accelerated the development of pre-trained language models in the general NLP domain (Wang et al., 2018, 2019; Rajpurkar et al., 2018; Lai et al., 2017) and biomedical applications (Peng et al., 2019; Gu et al., 2021). However, few comprehensive benchmarks exist for evaluating language models in the political conflict and violence domain. The focus of political science professionals is different from that of general NLP researchers. For example, they

[2] http://history-lab.org

are more interested in classifying, tracking, and predicting conflict events from the text.

To conduct a comprehensive evaluation of ConfliBERT, we collected a broad range of NLP tasks related to political conflict and violence from both publicly available and our newly-annotated datasets. Table 3 shows the datasets and their corresponding tasks. Some datasets may contain subsets and are related to various tasks. The table also lists the number of examples in the training, development, and test datasets as well as the evaluation metrics used for each task. In particular, we use F1 scores as performance metrics for binary classification tasks. We use example-based F1 metrics for multi-label classification tasks (Sorower, 2010). For all the other tasks, we rely on Macro F1 to assess the model's performance. Next, we describe the datasets and their tasks.

Binary classification (BC). We collected **BBC News** (Greene and Cunningham, 2006) and **20 Newsgroups** (Lang, 1995) for identifying political news, a subset from **Gun Violence Database** (Pavlick et al., 2016) for finding articles related to gun violence. We also used the samples from Global Contention Politics Dataset (**GLOCON**) (Hürriyetoğlu et al., 2019) to conduct one sentence-level and one document-level classification task to predict whether the story is related to protests. These BC tasks are essential for political scientists as a first step to classify and filter documents containing political and conflict events from large-scale news wires.

Multi-class classification (MCC). **GTD** refers to Global Terrorism Database which collects terrorist incidents from 1970 onward (START, 2019). We sampled a subset with description text longer than 40 words and single labels to classify 9 types of attacks such as bombing/explosion, armed assault, and hostage-taking.

India Police Events (Halterman et al., 2021) consists of sentences from English-language Times of India articles about police activity events in Gujarat during March 2002 (a relevant period due to widespread Hindu-Muslim violence). The labels are available for both document and sentence levels and consider five categories of police activity: kill, arrest, fail to act, force, and any action.

Event Status includes English news articles about civil unrest events annotated with temporal tags (Huang et al., 2016). Following the original setting, we conduct a temporal status classification (TS MCC) to detect the primary temporal distinctions among past, ongoing, and future. Besides, we also build a BC task of predicting if the story contains civil unrest events.

Multi-label classification (MLC). **SATP** stands for South Asia Terrorism Portal[3] from which we manually annotated a sample of 7,445 narratives between 2011 and 2019. We focus on incidents initiated by terrorist organizations. 23.6% of the sample are relevant stories classified into one or more categories: armed assault, bombing/explosion, kidnapping, and others. The rest samples are irrelevant (stories not about terrorism attacks such as arrests or armed clashes). Based upon this, we built three tasks. The first is a BC task to find relevant stories. The second is an MLC task to predict attack types on the relevant subset (Rel MLC). The third is the same as the second but conducted on the more imbalanced full dataset (All MLC).

InSight Crime (Parolin et al., 2021b) contains annotated stories about organized criminal activity in Latin America and the Caribbean from InSight Crime.[4] We applied an MLC task to predict multiple crime categories expressed in the stories, such as drug trafficking, corruption, and law enforcement.

Sequence Labeling or Named Entity Recognition (NER). **MUC-4** consists of documents reporting terrorism events, annotated with entities such as Perpetrator Individuals, Perpetrator Organizations, Physical targets, Victims, and Weapons (MUC-4, 1992). We split the dataset following Du and Cardie (2020).

Re3d stands for Relationship and Entity Extraction Evaluation Dataset (DSTL, 2018), comprising task-specific documents focused on the topic of the conflict in Syria and Iraq. The data contains annotations in span format with their corresponding entity types: Organization, Weapon, Military platform, Person, among others.

CAMEO (**C**onflict **a**nd **M**ediation **E**vent **O**bservations) is the industry standard for event extraction in political science (Gerner et al., 2002). An event classification, known as pentacode, consists of five event types: 0-Make a Statement, 1-Verbal Cooperation, 2-Material Cooperation, 3-Verbal Conflict, and 4-Material Conflict, and spans

[3] https://satp.org
[4] https://insightcrime.org

of texts containing sources (who conducted the action) and targets (to whom the action was conducted). We formulated two tasks for CAMEO event extraction on our newly-annotated dataset: sources and targets labeling (ST NER), and pentacode classification (PC MCC).

4 Experimental Setup

4.1 Pre-training Setup

We implemented ConfliBERT using two methods, Cont and SCR. Each approach has an uncased and a cased version. The architecture is the same as BERT-Base with 12 layers, 768 hidden units, 12 attention heads, and 110M parameters in total. Specifically, for our Cont models, we ran additional pre-training steps of the released checkpoints of BERT-Base models on our domain-specific corpus. The vocabulary is the same as the original BERT's vocabulary. For our SCR models, we use an in-domain vocabulary, ConfliVocab (See Section 3.1 for more details).

We discarded the next sentence prediction (NSP) task. We found that the predicted NSP accuracy quickly reached 90% in the middle of our training, which indicated that NSP might be less challenging for the model to learn in our domain. However, learning NSP simultaneously affected the speed of optimization of masked language models (MLM) loss. Following many recent works discarding NSP (Lample and Conneau, 2019; Yang et al., 2019; Joshi et al., 2020; Liu et al., 2019) and our observation, we optimized MLM only.

We used four V-100 GPUs with 32 GB memory to train each model. We used Adam optimizer (Kingma and Ba, 2015). The learning rate was warmed up over the first 10,000 steps to the peak value of 5e-4, and then linearly decayed. We pre-trained each SCR model for about 150K steps over the 7 billion word corpus. We followed Devlin et al. (2018) to train the model with a sequence length of 128 for 80% of the steps. Then, we trained the remaining 20% steps with a sequence of 512. The overall training time for each SCR model took about eight days. We trained Cont models the same as SCR models but in two fewer days because they were trained from intermediate checkpoints. See Appendix for more details.

4.2 Fine-Tuning Setup

Architecture. We followed the same architecture modification as BERT (Devlin et al., 2018) in the downstream tasks. Our task mainly consists of classification and sequence labeling. The sequence labeling tasks predict the sequence of BIO tags for each token in the input sentence. The classification tasks require a sequence classification/regression head on top of the pooled output of BERT. We used cross-entropy loss for binary/multi-class classification. We used mean-square loss and set the discrimination thresholds as 0.5 in all the multi-label classification tasks.

Casing. Devlin et al. (2018) use the cased models for NER and the uncased models for all other tasks. However, other works report that uncased models perform slightly better than cased models in specific domains, even on NER tasks (Beltagy et al., 2019; Gu et al., 2021). Therefore, we evaluated both cased and uncased versions of all models.

Hyperparameters. Devlin et al. (2018) propose a hyperparameter tuning strategy relying on a grid-search on the ranges such as the number of training epochs $\in \{3, 4\}$, and batch size $\in \{16, 32\}$. However, this strategy for general domain benchmarks (e.g. GLUE (Wang et al., 2018)) has not been sufficiently justified in other datasets (Chalkidis et al., 2020). The optimal hyperparameters are highly dataset- and task-dependent in our tasks. For instance, the models may be underfitting after the suggested maximum of four epochs. Additionally, based on our observations from the conflict datasets (e.g., GTD, SATP, MUC-4, InSight Crime), ConfliBERT models converge to the best results faster than BERT. Therefore, to compare with BERT fairly, we used early stopping based upon the development dataset within a range of the maximum training epochs when all the models have achieved stable results. A more detailed description of other hyperparameters can be found in the Appendix. Finally, we repeated all the experiments ten times with different seeds.

5 Results and Analysis

5.1 Pre-training Results

We use perplexity (ppl) to measure how well the language models predict a masked token in an unseen test set. We sampled 0.02% of stories from each source during the data preparation, ending with an 8.62 MB held-out dataset representing our corpus's distribution. Table 4 shows the ppl of our models on the held-out dataset. We also list the values reported by the original models (Devlin et al.,

Dataset	Domain	Tasks	Train/dev/test	Metrics	BERT		Confli.-Cont		Confli.-SCR	
					uncased	cased	uncased	cased	uncased	cased
BBC	General	BC	1588/315/322	F1	97.24	96.38	97.9	96.95	98.08	**98.13**
20 News.	General	BC	9044/2270/7532	F1	80.30	79.58	80.4	80.51	**81.05**	80.37
Gun V.	Violence	BC	3387/423/423	macro F1	84.30	85.24	90.02	**90.27**	86.35	86.13
GLOCON	Protest	Sent BC	1549/193/193	macro F1	84.53	84.92	85.60	85.72	**86.57**	82.20
		Doc BC	782/130/130	macro F1	88.97	84.61	89.76	89.97	**91.13**	88.27
GTD	Terrorism	MCC	2825/471/471	macro F1	83.55	82.05	81.97	83.23	**83.82**	83.16
SATP	Terrorism	BC	5956/744/745	F1	87.78	87.10	87.51	87.49	88.12	**88.72**
		Rel MLC	1085/232/232	example F1	87.81	88.36	88.14	88.37	**89.08**	88.64
		All MLC	4794/1192/1489	example F1	63.36	63.32	64.14	63.72	64.47	**64.53**
Insight C.	Crime	MLC	1002/332/319	example F1	68.57	67.83	69.09	69.15	68.68	**69.47**
India P.	Violence	Sent MLC	14943/3172/3276	example F1	64.89	64.54	63.03	63.40	**67.27**	66.22
		Doc MLC	905/165/187	example F1	66.80	63.41	67.09	67.38	**69.97**	66.71
Event S.	Protest	TS MCC	1818/226/227	macro F1	70.65	67.15	73.32	**75.03**	72.55	70.94
		BC	4010/500/501	F1	91.72	90.67	**92.42**	91.85	92.10	92.40
CAMEO	Politics	PC MCC	1348/224/225	macro F1	86.44	85.85	**87.88**	86.12	87.64	87.83
		ST NER	1153/224/225	macro F1	72.29	72.25	74.00	**74.45**	74.35	72.87
MUC-4	Terrorism	NER	1300/200/200	macro F1	62.96	60.33	60.29	60.90	**63.97**	60.31
Re3d	Defence	NER	574/191/200	macro F1	63.44	62.46	64.40	66.20	**66.40**	64.23

Table 3: The datasets, tasks and summary results of our evaluation.

	BERT	Confli.-SCR		Confli.-Cont	
	uncased	uncased	cased	uncased	cased
ppl	3.99	3.14	3.14	3.40	2.93

Table 4: Perplexity on held-out training data by model.

2018). Low ppl scores indicate that our models have been sufficiently pre-trained and have better generalization on our corpora.

5.2 Fine-Tuning Results and Analysis

Table 3 reports the F1 scores for each task using the mean of 10 seeds. We have the below observations:

ConfliBERT's superiority over BERT. ConfliBERT provides additional improvement to the original BERT in our target domain. In Table 3, although the performance is task-, dataset- and casing- dependent, our models consistently report the best results (in bold). In Figure 1, we compare ConfliBERT SCR-uncased with the best results from both cased and uncased versions of BERT in each experiment. We use different colors to denote four p-value thresholds ($p<0.01$, $p<0.05$, $p<0.1$, and $p\geq0.1$) of statistical significance. SCR-uncased demonstrates superior performance across all the tasks, and the difference is statistically significant at $p<0.1$ in all but three. Specifically for GTD, we observed that SCR-uncased slightly beats the best BERT, but it still shows a significant

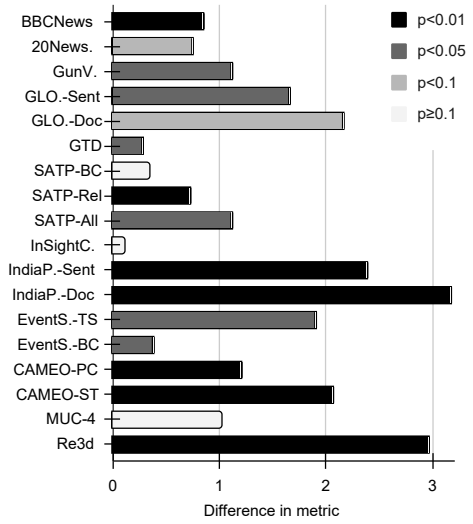
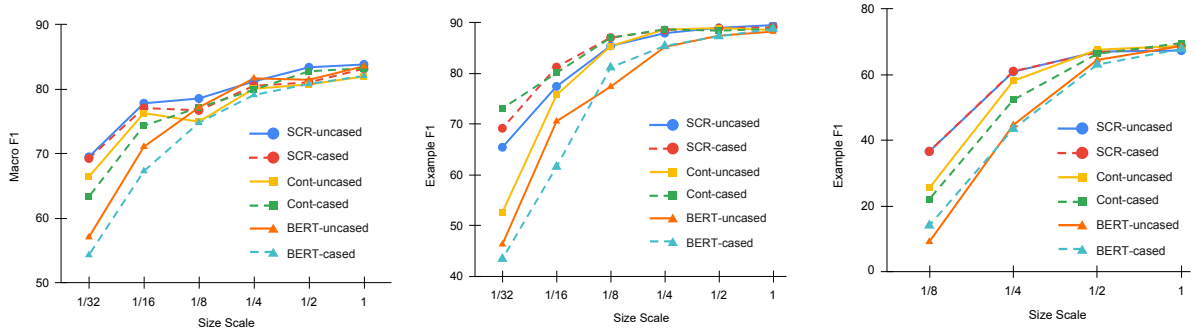


Figure 1: Significance test of SCR-uncased vs. the best of BERT models in each task, regardless of casing.

level of confidence, as depicted in Figure 1. We also observed that on InSight Crime, ConfliBERT achieves the best results in SCR-cased. Yet for SCR-uncased, the margin is not significant when compared with the best BERT in Figure 1. However, we conduct certain experiments on GTD and Insight Crime showing ConfliBERT's significant superiority when tackling limited training data in section 5.2.

Evaluating differences between the two pre-



(a) GTD (b) SATP-relevant (c) Insight Crime

Figure 2: Performance vs. varying size of the training data.

training strategies, Cont and SCR, remains to be studied. Table 3 shows that SCR slightly beats Cont in most cases (13 out of 18 tasks), and SCR-uncased provides the most stable improvements over BERT among our four models. However, the performance is still dataset- and task-dependent. For example, Cont beats SCR significantly in Gun violence and Event Status. We present an in-depth analysis of these two cases in Appendix.

Effect of ConfliVocab One major difference between SCR and Cont is the use of in-domain vocabulary. Section 5.2 shows that both Cont and SCR outperform BERT significantly while SCR slightly beats Cont. We have discussed the substantial advantage of using ConfliVocab during the tokenization in Section 3.1. Besides the examples in Table 2, in ConfliVocab we also find terrorist groups and criminal organizations frequently mentioned in the reports of violence and crime. Examples include Boko Haram, Al Qaeda, Sinaloa Cartel, PCC, FARC, Mara, among others. On the other hand, the range of actor entities in the politics domain is much larger and sparser than terrorist and criminal organizations. Given that we have a more distinct in-domain vocabulary in the conflict domain, we expect a more significant benefit from ConfliVocab in the conflict domain instead of the general politics domain.

ConfliBERT requires less annotated data than BERT. ConfliBERT performs well with limited data in various conflict datasets. Figure 2 shows three groups of experiments on GTD, SATP, and Insight Crime, where we used varying training data sizes but the same valid and testing set as the original experiments respectively. We repeated each experiment with five seeds and plotted the average metric scores.

Figure 2a shows that ConfliBERT beats BERT using limited size of GTD training data. Especially in the case of 1/32 size of GTD training data (88 examples), both SCR models still have 69% F1 scores, while BERT models drop to 55% F1 scores. In Figure 2b, we sampled various subsets of SATP-relevant, the SATP subset related to terrorist attacks. Results show that three of our models remain 65% to 73% F1 scores when using only 1/32 size of the training data (34 examples), while BERT drops to only 44% F1 scores. Finally, we also observe that both ConfliBERT SCR models significantly beat Cont and BERT models with a large margin on Insight Crime in Figure 2c.

These results show large improvements when using ConfliBERT with limited training data. Given the resources required to annotate data in conflict research, this is a particularly encouraging finding. These experiments also show that ConfliBERT outperforms BERT on GTD, SATP, and Insight Crime, strengthening the results in Figure 1.

6 Conclusion and Future Work

This paper presents the development, application, evaluation, and further exploration of ConfliBERT, a pre-trained language model for political conflict and violence. The development of ConfliBERT rests on an unprecedented effort on three fronts. First, we collect and curate a large domain-specific corpus to support the pre-training process. Second, we conduct a comprehensive evaluation across several datasets and various NLP tasks of distinct nature and varying degrees of complexity.

The results show that ConfliBERT consistently outperforms BERT in the conflict and political violence domain. Furthermore, the biggest improve-

ments are with limited training data, which conflict researchers often have due to the high costs of data annotation. In this way, ConfliBERT constitutes a valuable development that will contribute to a broad community of researchers in political science and policy sectors interested in tracking, analyzing, and predicting political violence and conflict on a global scale.

Due to limited time and computational resources, we did not conduct more experiments to explore various hyperparameters that could affect fine-tuning results, such as vocabulary size and pre-training epochs, to name a few. Future work should analyze how to optimize ConfliBERT, expand ConfliBERT to multi-lingual settings, and apply ConfliBERT to more challenging tasks such as understanding, inference, question answering, uncertainty qualification (Hu et al., 2021; Hu and Khan, 2021), and few-/ zero-shot tasks to speed up the study of NLP application for the political science community.

7 Ethical Impact

Our research considers several measures to mitigate concerns of bias in machine learning: (i) we implement standard social science practices to select corpora and training data (Barberá et al., 2021); (ii) for the pre-training stage, we gather a corpus with unprecedented global coverage to reduce regional biases; (iii) we move beyond the biases introduced from dictionary-based methods by using machine learning, as suggested by Wilkerson and Casas (2017); (iv) finally, we use multiple coders for the training data. However, copyright issues prevent us from sharing the raw data and hinder FAIR data principles (Wilkinson et al., 2016).

The broader goal of producing accurate and valid conflict data is to prevent or mitigate harm. These types of data provide a more objective means to understand and study conflict and armed violence. Our effort is an attempt to produce higher-quality data resources to serve this purpose.

Acknowledgments

The research reported herein was supported in part by NSF awards DMS-1737978, DGE-2039542, OAC-1828467, OAC-1931541, and DGE-1906630, ONR awards N00014-17-1-2995 and N00014-20-1-2738, Army Research Office Contract No. W911NF2110032 and IBM faculty award (Research).

References

Open Event Data Alliance. 2015. Petrarch: Python engine for text resolution and related coding hierarchy. http://www.github.com/openeventdata/petrarch.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, et al. 2018. Construction of the literature graph in semantic scholar. *NAACL HLT 2018*, pages 84–91.

Giuseppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

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

Pablo Barberá, Amber E Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler. 2021. Automated text classification of news articles: A practical guide. *Political Analysis*, 29(1):19–42.

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

John Beieler. 2016. Generating politically-relevant event data. *In Proceedings of the First Workshop on NLP and Computational Social Science*, pages 37–42.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.

Brian Blankenship. 2020. Promises under Pressure: Statements of Reassurance in US Alliances. *International Studies Quarterly*, 64(4):1017–1030.

Doug Bond, Joe Bond, Churl Oh, Craig J. Jenkins, and Charles 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.

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

Patrick T Brandt, Vito D'Orazio, Latifur Khan, Yi-Fan Li, Javier Osorio, and Marcus Sianan. 2022. Conflict forecasting with event data and spatio-temporal graph convolutional networks. *International Interactions*, pages 1–23.

Berfu Büyüköz, Ali Hürriyetoğlu, and Arzucan Özgür. 2020. Analyzing ELMo and DistilBERT on socio-political news classification. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 9–18, Marseille, France. European Language Resources Association (ELRA).

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Matthew J Connelly, Raymond Hicks, Robert Jervis, Arthur Spirling, and Clara H Suong. 2021. Diplomatic documents data for international relations: the freedom of information archive database. *Conflict Management and Peace Science*, 38(6):762–781.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

DSTL. 2018. Relationship and entity extraction evaluation dataset. https://github.com/dstl/re3d/. Accessed: 2021-07-01.

X. Du and Claire Cardie. 2020. Document-level event role filler extraction using multi-granularity contextualized encoding. In *ACL*.

Deborah J Gerner, Philip A Schrodt, Omür Yilmaz, and Rajaa Abu-Jabr. 2002. Conflict and mediation event observations (cameo): A new event data framework for the analysis of foreign policy interactions. *International Studies Association, New Orleans*.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017. Cross-lingual classification of topics in political texts. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 42–46, Vancouver, Canada. Association for Computational Linguistics.

Derek Greene and Pádraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proc. 23rd International Conference on Machine learning (ICML'06)*, pages 377–384. ACM Press.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

Andrew Halterman, Katherine Keith, Sheikh Sarwar, and Brendan O'Connor. 2021. Corpus-level evaluation for event qa: The indiapoliceevents corpus covering the 2002 gujarat violence. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4240–4253.

Andrew Halterman and Benjamin J. Radford. 2021. Few-shot upsampling for protest size detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3713–3720, Online. Association for Computational Linguistics.

Alex Hanna. 2017. Mpeds: Automating the generation of protest event data. Available at https://osf.io/preprints/socarxiv/xuqmv (2020/05/22). Unpublished Manuscript.

Yibo Hu and Latifur Khan. 2021. Uncertainty-aware reliable text classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 628–636.

Yibo Hu, Yuzhe Ou, Xujiang Zhao, Jin-Hee Cho, and Feng Chen. 2021. Multidimensional Uncertainty-Aware Evidential Neural Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7815–7822.

Ruihong Huang, Ignacio Cases, Dan Jurafsky, Cleo Condoravdi, and Ellen Riloff. 2016. Distinguishing past, on-going, and future events: The EventStatus corpus. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 44–54, Austin, Texas. Association for Computational Linguistics.

Ali Hürriyetoğlu, Hristo Tanev, Vanni Zavarella, Jakub Piskorski, Reyyan Yeniterzi, Deniz Yuret, and Aline Villavicencio. 2021. Challenges and applications of automated extraction of socio-political events from text (case 2021): Workshop and shared task report. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 1–9.

Ali Hürriyetoğlu, Erdem Yörük, Deniz Yüret, Çağrı Yoltar, Burak Gürel, Fırat Duruşan, and Osman Mutlu. 2019. A task set proposal for automatic protest information collection across multiple countries. In *European Conference on Information Retrieval*, pages 316–323. Springer.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *NeurIPS*.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin 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*, pages 146–157. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

J. Lu and Joydeep Roy. 2017. Universal petrarch: Language-agnostic political event coding using universal dependencies. Available at `https://github.com/openeventdata/UniversalPetrarch` (2020/05/22).

Roseanne W McManus. 2017. *Statements of Resolve: Achieving Coercive Credibility in International Conflict*. Cambridge University Press.

MUC-4. 1992. Fourth message understanding conference (muc-4). In *Fourth Message Uunderstanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*.

Clayton Norris, Philip Schrodt, and John Beieler. 2017. Petrarch2: Another event coding program. *Journal of Open Source Software*, 2(9):133.

Sean P. O'Brien. 2010. Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research. *International Studies Review*, 12(1):87–104.

Fredrik Olsson, Magnus Sahlgren, Fehmi ben Abdesslem, Ariel Ekgren, and Kristine Eck. 2020. Text categorization for conflict event annotation. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 19–25, Marseille, France. European Language Resources Association (ELRA).

Faik Kerem Örs, Süveyda Yeniterzi, and Reyyan Yeniterzi. 2020. Event clustering within news articles. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 63–68, Marseille, France. European Language Resources Association (ELRA).

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

Javier Osorio, Alejandro Reyes, Alejandro Beltrán, and Atal 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*, pages 49–56, Marseille, France. European Language Resources Association (ELRA).

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword fifth edition ldc2011t07. *Linguistic Data Consortium, Philadelphia*.

Erick Skorupa Parolin, Mohammadsaleh Hosseini, Yibo Hu, Latifur Khan, Javier Osorio, Patrick T Brandt, and Vito 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*.

Erick Skorupa Parolin, Yibo Hu, Latifur Khan, Javier Osorio, Patrick T Brandt, and Vito D'Orazio. 2021a. CoMe-KE: A New Transformers Based Approach for Knowledge Extraction in Conflict and Mediation Domain. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1449–1459. IEEE.

Erick Skorupa Parolin, Latifur Khan, Javier Osorio, Patrick Brandt, Vito D'Orazio, and Jennifer Holmes.

2021b. 3M-Transformers for Event Coding on Organized Crime Domain. *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*.

Ellie Pavlick, Heng Ji, Xiaoman Pan, and Chris Callison-Burch. 2016. The gun violence database: A new task and data set for nlp. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1018–1024.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65.

Alec Radford, Jeff Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Benjamin Radford. 2020a. Seeing the forest and the trees: Detection and cross-document coreference resolution of militarized interstate disputes. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 35–41, Marseille, France. European Language Resources Association (ELRA).

Benjamin J Radford. 2020b. Multitask models for supervised protests detection in texts. *arXiv preprint arXiv:2005.02954*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing acled: an armed conflict location and event dataset: special data feature. *Journal of peace research*, 47(5):651–660.

Sayeed Salam, Patrick Brandty, Jennifer Holmesy, and Latifur Khan. 2018. Distributed framework for political event coding in real-time. In *2018 2nd European Conference on Electrical Engineering and Computer Science (EECS)*, pages 266–273.

Philip A. Schrodt. 2006. Twenty Years of the Kansas Event Data System Project. *The Political Methodologist*, 14(1):2–6.

Philip A. Schrodt. 2009. TABARI. Textual Analysis by Augmented Replacement Instructions.

Erick Skorupa Parolin, Latifur Khan, Javier Osorio, Vito D'Orazio, Patrick T. Brandt, and Jennifer 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)*, pages 410–419.

Mohammad S Sorower. 2010. A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis*, 18:1–25.

START. 2019. The global terrorism database (gtd) [data file]. Retrieved from `https://www.start.umd.edu/gtd`.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Neural Information Processing Systems (NIPS)*, pages 5998–6008.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

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

John Wilkerson and Andreu Casas. 2017. Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, 20:529–544.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system:

Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja 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*, pages 19–27.

A Dataset

Table 5 and Table 6 list the detailed sources in our Expert Domain Corpora and Main Stream Media Collection described in Section 3.2, respectively.

Type: Sources	Size (MB)
United Nations: news.un.org, unodc.org, ohchr.org, unhcr.org, Refworld.org	427
U.S. Department of State: Annual Country Reports on Human Rights Practices, Annual Country Reports on Terrorism, International Religious Freedom Reports, Trafficking in Persons Report, Foreign Relations of the United States*	1,027
Non-government organizations: amnesty.org, hrw.org, rescue.org, phr.org, thenewhumanitarian.org, satp.org, cfr.org	838

*We filter a subset after World War II (Sep 1945 to 1989)

Table 5: Sources in Expert Domain Corpora.

Region	Sources	Size (GB)
Asia	Aljazeera, CNA, IndianTimes, JapanTimes, SCMP, TheNewsIntl, Xinhua	2.0
Europe	BBC, DW, France24, Guardian, Reuters, RFI, TASS	3.7
US	ABC, AP, CNBC, CNN, LATimes, NBC, NPR, NY Post, NYT, PBS, Politico, SFGATE, UPI, USA Today, US News, WASHPOST, WSJ	14.3
Others	AllAfrica, News24, EFE, TheConversation	0.8

Table 6: Sources in Mainstream Media Collection.

Filtering News Wires. We considered all the stories in EDC as relevant. However, for the general news in MMC, Gigaword, and PRT, we needed to filter our specific domain of political conflict and violence based on the websites' metadata information such as URLs, subjects, and tags. For example, we collected the stories with the tags such as Conflicts, Violence, War, Politics, Defense, Crime, et al. We also defined a bag-of-words classifier to assess unlabeled stories' relevance to our domain. Therefore, we statistically summarized two lists of the most frequent keywords' regular expressions from relevant stories and irrelevant stories. There are 266 patterns in the relevant list and 246 in the not relevant list. For example, our relevant list contains patterns such as `"activist"`, `"protest"`, `"counter.?terrorism"`, and `"jails?\b"`. Sports news use bellicose language similar to that of conflict stories with words such as `attack`, `shoot`, and `defeat`, thus presenting a classification challenge. The not relevant list contains frequent patterns such as `"shot \w+ goal"` to remove sports news. We compared the number of unique matching in the two lists and tuned the thresholds with the help of conflict experts. Finally, we filtered a small subset from MMC, Gigaword, and PRT in the conflict domain.

Filtering Wikipedia. We modified Wikiextractor (Attardi, 2015) to extract 18 GB size of documents with category labels from the Wikipedia dump[5] released on March 20, 2021. We used PetScan[6] to fetch pages of interest in the category hierarchy graph. We searched all the sub-categories within 0 to 4 depths under the union of five high-level topics: politics, activism, crime, government, and war. And we got 5 GB size of stories within 208,008 subcategories from the query. Then, we summarized the top 300+ frequent keywords from our targeted categories to prune irrelevant or too far-away child nodes based on the sub-category labels. We also removed unrelated categories such as fictional characters, movies, video war games, and historical events or people before the 20th Century, et al.

B Hyperparameters

Table 7 and Table 8 describe the detailed hyperparameters used in our pre-training and fine-tuning experiments, respectively. We implement our models using Huggingface API (Wolf et al., 2020).

[5] https://dumps.wikimedia.org
[6] https://petscan.wmflabs.org

Hyperparamter	SCR	Cont
Number of layers	12	12
Hidden Size	768	768
FFN inner hidden size	3072	3072
Attention heads	12	12
Mask percent	15	15
Learning Rate Decay	Linear	Linear
Warmup steps	10000	10000
Learning Rate LR	5e-4	5e-4
Adam ϵ	0.9	0.9
Adam β_1	0.98	0.98
Adam β_2	1e-6	1e-6
Attention Dropout	0.1	0.1
Dropout	0.1	0.1
Weight Decay	0.01	0.01
Train Steps	15,000	8,000
Vocabulary	ConfliVocab	BaseVocab
Uncased Vocab Size	30,000	30,552
Cased Vocab Size	30,000	28,996
Batch Size	2048	2048

Table 7: Hyperparamters for pre-training ConfliBERT using two strategies, pre-training from scratch (SCR) and continual pre-training (Cont). BaseVocab refers to the original BERT's vocabulary, while ConfliVocab refers to our domain-specific vocabulary.

Dataset - Tasks	max epochs	batch size	max seq-len	learning rate	drop-out
BBC News-BC	3	16	512	4e-5	0.1
20 News.-BC	3	16	512	4e-5	0.1
Gun V.-BC	10	8	512	5e-5	0.05
GLOCON-Sent BC	20	128	128	5e-5	0.05
GLOCON-Doc BC	5	8	512	5e-5	0.05
GTD-MCC	10	16	128	4e-5	0.1
SATP-BC	10	16	256	5e-5	0.05
SATP-Rel MLC	10	16	256	4e-5	0.1
SATP-All MLC	10	16	256	4e-5	0.1
InSight C.-MLC	5	16	512	4e-5	0.1
India P.-Sent MLC	10	16	128	4e-5	0.1
India P. - Doc MLC	10	16	512	4e-5	0.1
Event S.-TS MCC	10	192	150	5e-5	0.05
Event S.-BC	10	192	150	5e-5	0.05
CAMEO-PC MCC	40	32	128	5e-5	0.05
CAMEO-ST NER	60	32	128	5e-5	0.3
MUC4-NER	20	16	128	4e-5	0.1
Re3d-NER	25	16	128	4e-5	0.1

Table 8: Hyperparamters for fine-tuning all the models in our evaluation experiments.

C Other detailed results

This section analyzes in a detailed manner the model's performance on certain datasets. Specifically, we analyze two rare cases where all ConfliBERT models outperform BERTs and where Cont models significantly outperform SCR models. Table 9 indicates how Cont significantly outperforms SCR in all performance metrics ($p<0.05$ for all metrics). Table 10 shows how Cont-cased beats all the other counterparts for classifying event status of pieces of civil unrest. While there may be many factors, we postulate that some words in the original SCR-cased vocabulary are accidentally good at tokenizing the out-of-domain text in Gun Violence, while that vocabulary is also good at classifying ongoing (OG) and future (FU) events.

Tags	BERT		Confli.-Cont		Confli.-SCR	
	uncased	cased	uncased	cased	uncased	cased
0-TRUE	85.50	86.53	91.21	**91.39**	87.53	87.23
1-FALSE	83.11	83.95	88.84	**89.14**	85.17	85.02
Micro F1	84.40	85.36	90.17	**90.40**	86.47	86.23
Macro F1	84.30	85.24	90.02	**90.27**	86.35	86.13
AUROC	90.13	91.19	94.63	**95.45**	92.54	92.87
AUPRC	88.30	89.86	94.95	**95.76**	92.03	92.07

Table 9: Gun Violence Binary Classification.

Tags	BERT		Confli.-Cont		Confli.-SCR	
	uncased	cased	uncased	cased	uncased	cased
PA	88.38	87.28	89.27	**89.53**	88.92	88.63
OG	53.13	43.39	52.86	**56.17**	54.97	53.23
FU	70.45	70.77	77.82	**79.40**	73.76	70.94
Micro F1	79.47	77.49	81.28	**82.29**	80.48	79.87
Macro F1	70.65	67.15	73.32	**75.03**	72.55	70.94
MCC	56.67	51.45	60.40	**62.71**	59.74	57.40

Table 10: Event Status Temporal Status Classification.