ConfliBERT-Arabic: A Pre-trained Arabic Language Model for Politics, Conflicts and Violence

Sultan Alsarra[1], **Luay Abdeljaber**[1], **Wooseong Yang**[1], **Niamat Zawad**[1], **Latifur Khan**[1], **Patrick T. Brandt**[1], **Javier Osorio**[2], **Vito J. D'Orazio**[3]
[1]The University of Texas at Dallas, [2]The University of Arizona, [3]West Virginia University
{sultan.alsarra,luay.abdeljaber,wooseong.yang,
niamat.zawad,lkhan,pbrandt}@utdallas.edu,
josorio1@arizona.edu, vito.dorazio@mail.wvu.edu

Abstract

This study investigates the use of Natural Language Processing (NLP) methods to analyze politics, conflicts and violence in the Middle East using domain-specific pre-trained language models. We introduce Arabic text and present ConfliBERT-Arabic, a pre-trained language models that can efficiently analyze political, conflict and violence-related texts. Our technique hones a pre-trained model using a corpus of Arabic texts about regional politics and conflicts. Performance of our models is compared to baseline BERT models. Our findings show that the performance of NLP models for Middle Eastern politics and conflict analysis are enhanced by the use of domain-specific pre-trained local language models. This study offers political and conflict analysts, including policymakers, scholars, and practitioners new approaches and tools for deciphering the intricate dynamics of local politics and conflicts directly in Arabic.

1 Introduction

In the Middle East, political upheaval and carnage have long been issues (Blankenship, 2020). Deep divisions, geopolitical rivalries, and foreign meddling have historically riven the area, from the Israeli-Palestinian conflict to the ongoing civil war in Syria. Even if the root causes of these conflicts are complex and multidimensional, the role that language and communication play in shaping the narratives that underlay them cannot be ignored. Language is commonly used as a strategy to rally support, defend violence, and discredit opposing viewpoints. Therefore, it is essential to develop effective methods for understanding and analyzing the role that language and texts plays in Middle Eastern politics and conflicts via news reports and other sources. Natural Language Processing (NLP) approaches can evaluate large amounts of text and have shown great promise in identifying patterns and insights that would otherwise be difficult to spot. Recent, pre-trained language models (PLM), like BERT (Devlin et al., 2018), have improved in efficiency for a range of NLP tasks, including sentiment analysis, text categorization, and language synthesis. PLMs have received a lot of attention in the literature, but most of it has focused on English or other widely spoken languages; very few studies have examined how well they apply to Arabic. The Arabic language has a rich and complicated morphology, which has increased the requirement for highly advanced NLP tools that can meet the language's expanding needs across a variety of fields and applications (Ameur et al., 2020).

This research fills this vacuum in the literature by investigating the application of Arabic-specific PLMs for politics, conflicts and violence in the Middle East. We reference a variety of pertinent academic works, such as investigations into the nature of political violence (Asal et al., 2020), the function of language in conflicts (Webber et al., 2020), and the creation and use of PLMs (Jawahar et al., 2019; Devlin et al., 2018) such as ConfliBERT (Hu et al., 2022).

The performance of two PLMs, BERT and ConfliBERT-Arabic, focuses on the analysis of Arabic text about politics, conflict, and violence in the Middle East. BERT is a more general-purpose PLM that has been used to tackle a variety of NLP problems, whereas ConfliBERT-Arabic is a domain-specific PLM optimized on a corpus gathering texts relevant to regional politics, conflicts and violence. We contrast their effectiveness with a different PLM, AraBERT (Antoun et al., 2020).

This work has implications for multiple users such as policymakers, researchers, and conflict analysts. By providing cutting-edge tools and methods for investigating politics and conflicts in the Middle East, our study develops data for more effective conflict prevention and resolution programs. By

Proceedings of Recent Advances in Natural Language Processing, pages 98–108
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_011

examining the role that language and communication play in affecting the politics and conflicts in the region, we can provide a more nuanced understanding and prediction of the underlying causes of these conflicts and cooperation in the Middle East.

Our experiments show that domain-specific pretraining significantly improves model performance, particularly for identifying information about political conflict. We examine in detail each model's applications and their benefits and drawbacks.

2 Challenges

2.1 The Arabic Language

The Arabic language possesses distinctive characteristics that set it apart from English. A single Arabic character can take up to three different forms, each corresponding to a specific position within a word (beginning, middle or end). Moreover, Arabic lacks capital letters, which poses a considerable challenge for NER tasks, where capitalization plays a crucial role in other languages (Alkhatib et al., 2020). Arabic also has long and short vowels, but short vowels are no longer used in newspapers, leading to high ambiguity in texts as disambiguation using these vowels is impossible. In word disambiguation in Arabic, the diverse pronunciations of a word can give rise to various meanings. These small signs added to letters help readers differentiate between similar words. Nonetheless, omitting diacritics from some words can result in numerous lexical ambiguities (Laatar et al., 2018). Lastly, Arabic is highly inflectional, with a very complex morphology. The general form of an Arabic word comprises Prefix(es) + Stem + Suffix(es), with the number of prefixes and suffixes ranging from 0 or more. Affixes are added to the stem to obtain the required expression. For example, the Arabic word "manzil" means "house," while "almanzil" means "the house," illustrating how an Arabic word can be translated into two words. Another example, "sayaktoubounaha," which means "and they will write it," when written in the general form introduced above, becomes sa+ya+"ktoub"+ouna+ha. From an NER perspective, this peculiarity of Arabic poses a significant obstacle as it results in data sparseness (Benajiba et al., 2007).

2.2 Corpora Building

When scraping Arabic sites, text encodings must be in UTF-8 for the text to be processed by NLP. This also accounts for the Arabic text direction, from right to left, and proper encoding ensures that this feature is recognized (Meskaldji et al., 2018). Several technical issues are that, 1) Arabic sites store limited data due to high database costs; 2) Security features on many Arabic sites can hinder scraping efforts. Thus trial, and error runs are necessary to determine the optimal number of parallel threads and sleep time between consecutive scrapes of the relevant sites. Since some Arabic websites present static news stories on individual pages while others generate dynamic stories, scripts had to be written from scratch, tailored to the structures of individual news websites. Finally, it was essential to ensure that the sites being scraped are written in modern standard Arabic (MSA).

3 Related Work

Recent developments in pre-trained language models have significantly advanced the field of Natural Language Processing. Here, we review three of the most prominent models: BERT, Multilingual BERT, AraBERT, and ConfliBERT.

3.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) is a Google developed PLM (Devlin et al., 2018). BERT is trained on a massive corpus of text using an unsupervised learning method that involves predicting missing words in a sentence. BERT has demonstrated superior performance on various Natural language Processing tasks, including sentiment analysis, question answering, and language translation. To fine-tune BERT for specific tasks, a task-specific layer is added on top of the pre-trained model, and the whole architecture is trained on a labeled dataset. This approach has shown to achieve state-of-the-art results in several Natural Language Processing tasks. However, one of the limitations of BERT is its focus on the English language.

3.2 Multilingual BERT

Multilingual BERT is an improved version of BERT that addresses the language dominance of the original model (Pires et al., 2019). Multilingual BERT outperforms the original BERT in several languages. For tokenization, Multilingual BERT uses a 110k shared WordPiece vocabulary file that spans 102 languages. Similar to the English BERT, lower casing+accent removal, punctuation splitting, and whitespace tokenization were applied. How-

ever, the Arabic language's complex concatenative (Al-Sallab et al., 2017) system poses a challenge for BERT-compatible tokenization, as words can have different forms but the same meaning. Therefore, when using BERT-compatible tokenization, tokens appear twice, once with the Arabic definite article "ال" (equivalent to "the" in English) and once without it, leading to unnecessary redundancy (Antoun et al., 2020).

3.3 AraBERT

AraBERT is a PLM specifically designed for Arabic language understanding (Antoun et al., 2020). AraBERT is trained on a large Arabic corpus using the same methodology as BERT, but uses different tokenization. The authors segment words using Farasa (Abdelali et al., 2016) into stems, prefixes, and suffixes, and then train a SentencePiece, an unsupervised text tokenizer and detokenizer, in unigram mode on the segmented pre-training dataset to produce a subword vocabulary of approximately 60k tokens. One of the limitations of AraBERT is that the training corpus is not domain-specific, compiled from Arabic Wikipedia and other public datasets.

3.4 ConfliBERT

ConfliBERT is an English PLM designed for conflict and political violence (Hu et al., 2022). ConfliBERT is trained on a large domain-specific corpus using a multi-task learning method to perform several related tasks simultaneously. ConfliBERT has demonstrated superior performance on several political violence detection tasks with external validation (Häffner et al., 2023). ConfliBERT is fine-tuned with a task-specific layer added on top of the PLM, and the entire architecture is trained on a labeled dataset for the downstream task. ConfliBERT has been expanded to Spanish with the introduction of ConfliBERT-Spanish (Yang et al., 2023)

Overall, each model has its strengths and limitations. While BERT and its variants have proven to be effective in several NLP tasks, they have a limited focus on the Arabic language. In contrast, AraBERT is specifically designed for Arabic language understanding, but its training corpus is not domain-specific. Our work aims to build upon the strengths of previous language models to create a specialized model that is tailored to the Arabic language and the domain of political violence. By combining the features and methodologies of BERT, AraBERT, and ConfliBERT, we aim to develop a model capable of accurately detecting and analyzing instances of political conflicts and violence in Arabic texts.

4 Approach

To develop ConfliBERT-Arabic, we implemented a series of steps, namely corpora collection, pre-training strategies, and evaluation tasks. The first step involves the creation of a domain-specific corpus for pre-training. Publicly available Arabic datasets focusing on politics and conflict domains are limited, and thus we conducted our own data collecting to extract political text from Arabic sources, thus enabling us to achieve better results on political tasks. After building the corpus, we developed our domain-specific model based on BERT, a powerful language model that has been successfully validated in multiple domains in both English and Arabic languages (Lee et al., 2019; Beltagy et al., 2019; Chalkidis et al., 2020; Gu et al., 2021; AL-Qurishi et al., 2022; Bayrak and Issifu, 2022; Boudjellal et al., 2021). Our Masked-Language Modeling (MLM)-based BERT model shows improved performance compared to other transformer models that use different self-supervision tasks. The final step involves evaluating the performance of ConfliBERT-Arabic on downstream tasks related to political and conflict analysis to measure its effectiveness in real-world applications.

4.1 Corpora Building

During our data collection process, we scraped a total of 84 sources from various Arabic language speaking countries. A total of 19 countries were covered in the corpus building. These sources consisted of newspaper sites, mainstream media, and government sources such as national news agencies. The list of sources were curated and scraped by native Arabic speakers to ensure all sources were in Modern Standard Arabic (MSA). Our focus during scraping was on news from the Political, International, and Local sections of the sources, as we determined that these categories provided a greater proportion of political, conflict and violence-based articles. To ensure the highest quality of data, we ignored sections focusing on Culture, Entertainment, Economy, Business, and Sports. In total, we were able to extract 11.5GB of data from these sources. To construct the corpora, we followed these steps:

1. For several Arab countries, we curated a list of official national news agencies. We also included national news agencies of countries with Arabic as a second language such as Mauritania, Kyrgyzstan and Tajikistan.

2. We curated a list of newspapers that are widely circulated and considered reliable sources for news. We focused on highly political countries in the region which are Palestine, Saudi Arabia, Lebanon, Syria, and Iraq.

3. We curated a list of well known Mainstream Media Sources in the region such as BBC Arabia and Aljazeera. A few of these resources were run by non-arab countries, but targeted the region with arabic sites, such as Russia Today Arabia by Russia and Adnki by Italy.

4. We created python scripts to extract text from the list of sources using high performance computers (HPC) that have 96 cores with 10 GB memory each.

5. For each source, we processed and cleaned the data. This involved removing duplicate texts, carriage returns, peripheral punctuation marks, extra white space, and pop-up advertisements text.

6. We stored the extracted data in a CSV files using arabic friendly UTF-8 encoding. The CSV includes metadata such as country name, outlet name, article title, link, and date.

After scraping the data, we designed a filtering technique to reduce the possibility of irrelevant news articles. For example, the international section of a newspaper might include a story about an Olympic Games match between two politically rival countries. While this article may have political implications for the countries involved, we consider it more relevant to sports than to our domain. To create our filter, we built a list of relevant and irrelevant keyword. The keywords were created after verbs and actors in the CAMEO dictionary (Gerner et al., 2002) and reviewed by experts in the political science domain. The number of matches with the relevant and irrelevant keywords were compared against each other and the thresholds was tuned to filter the most relevant political, violence and conflict-based news. Table 2 shows statistics of the extracted corpora after filtering.

4.2 Domain-Specific Pretraining

As shown in Figure 1, we employed a continual method (Cont) to adapt BERT to the political, conflict, and violence domain. This method involves initializing the BERT's model vocabulary and checkpoints, then training the model for additional steps on our domain-specific corpus. We used Multilingual BERT and AraBERT as the base BERT models for our Cont method. Since Multilingual BERT and AraBERT have already been pretrained about one million steps on a generic arabic domain, the Cont method will require fewer steps than training from scratch. The Cont method has shown comparable results to training from scratch. According to (Lee et al., 2019), continuous pretraining of BERT on a biomedical dataset for 470K steps results in performance comparable to pretraining for one million steps.

When it comes to casing, although there is no distinction between upper and lowercase letters in Arabic, previous works for English (Beltagy et al., 2019; Gu et al., 2021; Devlin et al., 2018) have shown, in specific domains, uncased models perform slightly better than cased models especially when it comes to NER tasks. Therefore, we decided to evaluate both cased and uncased versions of Multilingual BERT for Arabic to highlight any differences and to be comprehensive in our research.

4.3 Evaluation Tasks

The development of pre-trained language models has been accelerated by the introduction of comprehensive benchmarks in the general NLP domain (Wang et al., 2018, 2019; Rajpurkar et al., 2018; Lai et al., 2017), as well as in biomedical applications (Peng et al., 2019; Gu et al., 2021). To comprehensively evaluate ConfliBERT-Arabic, we collected a diverse set of datasets for Named Entity Recognition (NER) and Binary Classification (BC). However, we faced a challenge as we could not find any comprehensive benchmarks for evaluating Arabic language models specifically in the political, conflict and violence domain.

The focus of Arabic NLP in recent research has mainly been on social media and dialect detection. Luckily, we did find some news-based datasets, but they covered a wide range of news topics which included politics. Therefore, we had to filter these news based datasets to isolate the political, conflict and violence related sections. As for datasets

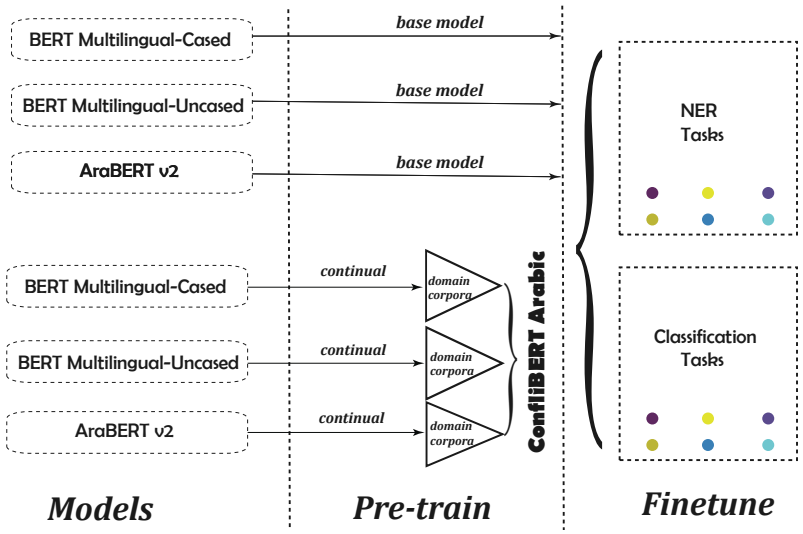

Figure 1: Workflow of our ConfliBERT-Arabic framework.

Country	Source	Size (MB)	Country	Source	Size (MB)
Palestine	Alquds Alarabi Newspaper	1169	**Syria**	Syrian Arab News Agency	208
	Alsbah Newspaper	2.8		Al-Ba'ath	243
	Sonara Newspaper	4		Enab Baladi	174
	Donia Alwatan Newspaper	1220		Aks alser	209
	Alresalah Newspaper	310		Aljaml	543
	Mashreq News	241		Orient News	199
	Al-Meezan Newspaper	1.35		Al-Wehda	5.31
	Felesteen News	164		Syrian Network for Human Rights	5.08
Saudi Arabia	Asharq Al Awsat Newspaper	8.04		Al-Ourba	23.6
	Al-Jazirah Newspaper	68		Al-Furat	19.3
	Albilad Newspaper	149		Al-Fedaa	24
	Majalla Magazine	38.8		Syrian News Station	357
	Makkah Newspaper	187		Al-Ghad	859
	Al Watan Newspaper	18.2	**Iraq**	National Iraqi News Agency (NINA)	316
	Al arabiya	282		Almustakbal Paper	88.6
	Alhadath	114		Al Sabah Newspaper	41.6
	Sra7h	188		Al Sabah Al Jadeed	5.36
	Rwifd	16.9		Kitabat Newspaper	102
Lebanon	National News Agency	322		Aladala news	7.2
	Al Joumhouria Newspaper	445.5		Alaalem newspaper	8.52
	Aliwaa News Paper	140		Alsumaria	37.4
	Elnashra	32.7		Al basrah Paper	30.9
	Albadeel	1		Almuraqeb Aliraqi	69.4
	Al-Binaa	475.2		Tareeq Ashaab	4.9
	Bintjbeil	185		Alnahda	2.14
	Lebanon 24	19.4		Basra Press 24	4.51
	Kataeb.org	8.22		Alfurat	209
	Janoubia	258		Bader News	1.21
	Al-Ahed	156.4		Azzaman	364
	Cedar News	188		Alrasheed	79.8
	Almayadeen	6.16	**Egypt**	Arabnet5	138
	Ch 23	53.3	**Sudan**	Sudan News Agency	0.228
	Murr TV (MTV)	2.66	**Libya**	Jamahiriya News Agency	22.7
	LBC	0.245	**Qatar**	Aljazeera	69.7
	Saida TV	9.48	**Morocco**	Attarik	2.58
	Inbaa	57.9	**UK**	Qudspress	3.86
USA	CNN Arabia	17.9		BBC Arabia	12.2
Turkey	TR Agency	31.5		Elaph	251
Kyrgyzstan	Kyrgyz National News Agency	4.38	**Tajikistan**	National Information Agency of Tajikistan	5.92
Russia	Russia Today Arabia	4.46	**Mauritania**	Mauritanian News Agency	6.58
	Sputniknews Arabia	109		AMP	5.31
Iran	Alalam	27.1	**Italy**	Adnki	46.5

Table 1: List of sources used for corpora

Parameters	Count
Number of Articles	2,995,874
Words (Tokens)	928,729,513
Number of Sentences	24,221,481
Average Number of Words Per Sentence	38.34
Overall Token Frequency	5,924,007

Table 2: Statistics of the extracted political and conflict-specific dataset

that focused on Wikipedia and social media, although these datasets may not cover the political domain specifically, we decided to evaluate them using our model. The reason is many of these datasets did include political posts from social media and Wikipedia that reference conflicts and violence. Moreover, we noticed social media in the MENA region is highly political and full of conflicts and violence, and we were interested in the results. Below we present the datasets and their corresponding tasks.

4.3.1 Binary Classification (BC)

Political scientists need Binary Classification to identify political and conflict-related documents from large news corpora. We gathered several datasets, including **SANAD** (Einea et al., 2019), **Ultimate Arabic News** (Al-Dulaimi, 2022), **AraFacts**(Ali et al., 2021), **DataSet for Arabic Classification**(mohamed, 2018) and **Arabic Dialects and Topics** (Boujou et al., 2021). SANAD comes from AlArabiya, Akhbarona, and Alanba AlKhaleej newspaper websites. We created two BC tasks here one for AlArabiya and one for Alanba AlKhaleej. Ultimate Arabic News is a collection of single-labeled texts from Arabic news websites and press articles. AraFacts contains claims from five Arabic fact-checking websites, mostly of political nature. DataSet for Arabic Classification consists of 111,728 documents collected from the Arabic online newspapers Assabah, Hespress and Akhbarona. Finally, we acquired Arabic Dialects and Topics, which is a dataset for topic detection for social media posts in different Arabic dialects. These datasets cover a wide range of text types, but we focused on evaluating their performance for Binary Classification tasks.

4.3.2 Named Entity Recognition (NER)

The NER datasets we selected are annotated in CoNLL format and contain entities such as location, organization, person, group, event, and others. The datasets are as follows: **KALIMAT** (El-Haj and Koulali, 2013), which includes documents from the Omani newspaper Alwatan; **AnerCORP** (Benajiba et al., 2007), a publicly available Arabic NER dataset from news sources with 150,286 tokens and 32,114 types; **AQMAR** (Mohit et al., 2012), which is a corpus of 74,000 tokens from 28 annotated Arabic Wikipedia articles; **Wikiann** (Rahimi et al., 2019), a manually annotated dataset covering approximately 3,000 sentences from 31 Wikipedia articles; **LinCE MSA-EGY** (Aguilar et al., 2019), an annotated social media dataset using Twitter, where the tweets were harvested from the timeline of 12 Egyptian politician public figures; **WDC** (Althobaiti et al., 2014), which contains 165,119 sentences from Wikipedia, consisting of around 6 million tokens; and finally, **POLYGLOT-NER** (Al-Rfou et al., 2015), a generated annotated dataset using Wikipedia and Freebase.

5 Experimental Setup

5.1 Pre-training Setup

We implemented ConfliBERT-Arabic using the previously mentioned continual (Cont) techniques. The architecture used is similar to Multilingual BERT-Base with 12 layers, 768 hidden units, 12 attention heads, and a total of 110M parameters. The vocabulary file used is identical to the original Multilingual BERT and AraBERT vocabulary files. We used 2 Nvidia A-100 GPUs with 10 GB memory to train the models. We used an Adam optimizer (Kingma and Ba, 2015) with the learning rate set to a peak value of 5e-5 and then linearly decayed. To accommodate the long paragraphs of new data, we trained the model with a sequence length of 512. The overall training time for each Cont model took about three days.

5.2 Fine-Tuning Setup

For Named Entity Recognition (NER) tasks, we predicted the sequence of BIO tags (a common tagging format for tagging tokens in a chunking task) for each token in the input sentence. We pre-processed the dataset to ensure the correct CoNLL format and used a (70,15,15) split for Train, Test, and Dev for all datasets. For Binary Classification (BC), we required a sequence classification/regression head on top of the pooled output of BERT. We used cross-entropy loss for binary classification. We split our datasets into (70,15,15)

Model		NER F1 Score	BC F1 Score
ConfliBERT Arabic	Multilingual-Uncased-Cont	**77.07**	**90.85**
	Multilingual-Cased-Cont	**77.14**	**90.78**
	AraBERT-Cont	**77.88**	**91.54**
BERT multilingual	Uncased	76.69	89.12
	Cased	76.86	89.10
AraBERT		75.89	90.16

Table 3: Summary F1 results of our evaluation by task

Dataset	Domain	ConfliBERT-Arabic F1 Score			BERT F1 Score		
		AraBERT	Cased	Uncased	AraBERT	Cased	Uncased
AnerCORP	Newswire	**81.17**	77.74	77.75	79.7	75.23	75.46
Kalimat	Newspaper	82.09	**83.72**	82.37	81.53	82.74	82.63
LinCE	Social Media	**79.96**	79.19	79.67	77.47	76.39	76.59
WikiANN	Wikipedia	**92.97**	92.06	92.2	92.88	91.73	91.68
WDC	Wikipedia	72.91	72.85	72.72	71.49	**73.03**	73.27
Polyglot	Wikipedia	**64.61**	62.48	62.35	60.111	62.66	62.03
Aqmar	Wikipedia	71.45	71.95	72.4	68.07	**76.28**	75.23

Table 4: Summary of F1 measure results of NER datasets

for Train, Test, and Dev. We fine-tuned our models on a single Nvidia A-100 GPU for five epochs with a learning rate of 5e-05, batch size of 16, and a maximum sequence length of 128 for NER and 512 for BC. We repeated all experiments ten times with different seeds. We use F1 scores as performance metrics for both tasks.

6 Results and Analysis

Table 3 reports the F1 scores for each model by task with results using the mean of 10 seeds. As shown, ConfliBERT-Arabic outperforms Multilingual BERT and AraBERT, where our models consistently report the best results (in bold) for both tasks. To compare ConfliBERT-Arabic Continual with other models, we evaluated the best results from cased, uncased, and AraBERT versions of BERT. Our findings show that ConfliBERT-Arabic Continual based on AraBERT performs the best overall by achieving the top results in 9 out of the 13 datasets evaluated. Overall, the models fine-tuned on our data had the best results in 11 out of the 13 datasets.

6.1 NER Evaluation Results

In Table 4, we can observe that our models outperformed the competing models on 5 out of the 7 evaluated datasets. Notably, our models demonstrated significant improvements across various types of datasets, including news articles, Wikipedia entries, and social media content, particularly when the datasets involved topics related to politics and international affairs.

In news-based datasets such as AnerCORP and Kalimat, our continual models demonstrated improvements over standard BERT. AnerCORP contained a significant amount of political and international data, with 34.8% of the dataset originating from Aljazeera.net, which primarily featured political articles focusing on conflict and violence. Consequently, our continual models exhibited considerable enhancements compared to standard BERT models. Similarly, for Kalimat, which was collected from the Omani newspaper Al-Watan, our models performed better, as the dataset mainly consisted of local and international news that covered the gulf region's political conflicts.

Regarding LinCE, researchers focused on social media data obtained from Twitter, specifically 12,334 tweets posted by 12 Egyptian political public figures. As the dataset was predominantly political discussing the region conflicts and featured numerous political named entities, our models outperformed standard BERT models.

For Wikipedia-based datasets, our performance varied depending on the specific articles used in each dataset. In the case of Polyglot, our models excelled due to the political and conflict/war oriented Wikipedia articles extracted using Freebase. Similarly, WikiANN, which contained political and conflict-related articles, led to our models performing well.

On the other hand, Aqmar and WDC, which consisted of more general articles unrelated to politics or conflict, we witnessed a better performance from the regular BERT models. For instance, Aqmar included 28 Wikipedia articles covering history, science, sports, and technology, as researchers aimed to adapt named entity analysis to new domains. In the case of WDC, the articles were sourced from Wikipedia's open domain, representing various genres. Here, the baseline cased multilingual BERT marginally outperformed our models. This was expected, as our models were pretrained and specialized for the political and conflicts domain.

For the NER datasets, we used p-values to confirm the statistical significance of the differences. Using AnerCORP, LinCE, and Polyglot, we contrasted our models with AraBERT and Multilingual BERT. All results are statistically significant at $p<0.01$. This makes sense given that all of these datasets have a strong political focus. In contrast, generic datasets such as Aqmar, had a $p>0.1$.

Dataset	Domain	ConfliBERT-Arabic F1 Score			BERT F1 Score		
		AraBERT	Cased	Uncased	AraBERT	Cased	Uncased
Ultimate Arabic News	News	**97.46**	95.85	95.89	95.74	94.27	94.80
DataSet for Arabic Classification	News	**97.47**	97.01	97.21	97.05	96.15	96.18
Arabic Dialects & Topics	Social Media	**67.09**	60.87	60.93	60.01	59.90	60.40
SANAD	AlArabiya News	**98.83**	97.81	98.01	98.42	97.11	97.13
	AlKhaleej News	**99.51**	99.09	99.07	98.93	98.02	98.22
Arafacts	Fact Checking	**75.21**	72.83	72.57	72.02	70.34	67.55

Table 5: Summary of F1 measure results for classification dataset

Given that our models was trained on a corpus of political domain data, it makes sense.

In summary, our models demonstrated superior results when applied to datasets rich in political, international, and conflict-related content, regardless of whether the data was sourced from news outlets, Wikipedia, or social media. For datasets that do not involve these topics, regular BERT models tended to yield better results.

6.2 BC Evaluation Results

Binary classification results are illustrated in Table 5. All datasets exhibited improved performance with our models. These datasets included four from newspapers, one from social media, and one from a fact-checking site. Since the datasets were originally created for topic classification purposes, all articles were annotated and labeled by categories such as Culture, Finance, Medical, Politics, Religion, Sports, and Tech.

To create our binary classification dataset, we sampled articles from the politics category, with emphasis on conflict and violence, alongside non-political data from other topics. The data was then labeled with 0 or 1 to indicate whether the articles were related to political conflict and violence or not.

Our ConfliBERT Arabic Continual model, based on AraBERT, demonstrated the best performance across all datasets except for one, where our uncased version performed better. Additionally, our models performed exceptionally well on political tweets and a fact-checking site, which also featured a significant amount of political content.

Again, we used p-values to confirm the statistical significance of the differences for datasets including Arabic Classification, SANAD Alarabiya, and SANAD AlKhaleej where there were only marginal gains in F measure. We contrasted the best outcomes from our models with the baseline iterations of BERT in each experiment. In all three tasks, our models performed better, with statistical significance set at $p<0.01$.

7 Conclusion and Future Work

In this paper, we introduce ConfliBERT-Arabic, a pre-trained language model for politics, conflict and violence in Arabic-language. ConfliBERT-Arabic's development required the acquisition and curation of a sizable domain-specific corpus for the pre-training stage. We also thoroughly assessed the model's performance across a range of NLP tasks and datasets, showing that ConfliBERT-Arabic regularly outperforms BERT in the politics, conflict and violence domain, especially when working with sparse training data. Researchers and decision-makers interested in tracking, analyzing, and predicting political violence and war in the Middle East will find these findings to be of great interest. ConfliBERT-Arabic is an important advancement that will help a large community of political scientists and decision-makers as a whole.

In future work, we are planning to expand on parameters such as vocabulary size and epochs to better optimize ConfliBERT-Arabic. Additionally, applying ConfliBERT-Arabic to more challenging tasks such as understanding, inference, question answering, and uncertainty qualification is planned

Acknowledgments

This research was supported in part by NSF awards DGE-2039542, OAC-1828467, OAC-1931541, and DGE-1906630, ONR award N00014-20-1-2738, Army Research Office Contract No. W911NF2110032 and IBM faculty award (Research). We thank the High Performance Computing (HPC) resources supported by the University of Arizona TRIF, UITS, and Research, Innovation, and Impact (RII), and maintained by the UArizona Research Technologies department. This work used Delta GPU at the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign through allocation CIS220162 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants 2138259, 2138286, 2138307, 2137603, and 2138296.

Ethical Impact

To address concerns of bias in machine learning, our research employs several measures. Firstly, we utilize standard social science practices to select corpora and training data (Barberá et al., 2021). Secondly, we gather a corpus with unparalleled global coverage for the pre-training stage, which aims to reduce regional biases. Thirdly, we move beyond biases inherent in dictionary-based methods by utilizing machine learning techniques, as suggested by Wilkerson in (Wilkerson and Casas, 2017). Lastly, we use multiple coders for the training data. However, due to copyright issues, we are unable to share the raw data, which hinders the principles of FAIR data (Wilkinson et al., 2016). The overarching aim of our research is to generate accurate and reliable conflict data to prevent or mitigate harm. These data offer an objective means to comprehend and examine conflict and armed violence. Our research endeavors to produce superior data resources to fulfill this purpose.

References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for arabic. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations*, pages 11–16.

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2019. Named entity recognition on code-switched data: Overview of the calcs 2018 shared task. *arXiv preprint arXiv:1906.04138*.

Ahmed Hashim Al-Dulaimi. 2022. Ultimate arabic news dataset.

Muhammad AL-Qurishi, Sarah AlQaseemi, and Riad Soussi. 2022. Aralegal-bert: A pretrained language model for arabic legal text.

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Polyglot-ner: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 586–594. SIAM.

Ahmad Al-Sallab, Ramy Baly, Hazem Hajj, Khaled Bashir Shaban, Wassim El-Hajj, and Gilbert Badaro. 2017. Aroma: A recursive deep learning model for opinion mining in arabic as a low resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 16(4):1–20.

Zien Sheikh Ali, Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. 2021. Arafacts: the first large arabic dataset of naturally occurring claims. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 231–236.

Manar Alkhatib, Azza Abdel Monem, and Khaled Shaalan. 2020. Deep learning for arabic error detection and correction. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(5):1–13.

Maha Althobaiti, Udo Kruschwitz, and Massimo Poesio. 2014. Automatic creation of arabic named entity annotated corpus using wikipedia. In *EACL 2014-14th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Student Research Workshop*, pages 106–115. The Association for Computer Linguistics.

Mohamed Seghir Hadj Ameur, Farid Meziane, and Ahmed Guessoum. 2020. Arabic machine translation: A survey of the latest trends and challenges. *Computer Science Review*, 38:100305.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. arxiv 2020. *arXiv preprint arXiv:2003.00104*.

Victor Asal, Carter Johnson, and Jonathan Wilkenfeld. 2020. Ethnopolitical violence and terrorism in the middle east. In *Peace and conflict 2008*, pages 55–66. Routledge.

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.

Giyaseddin Bayrak and Abdul Majeed Issifu. 2022. Domain-adapted bert-based models for nuanced arabic dialect identification and tweet sentiment analysis. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 425–430.

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.

Yassine Benajiba, Paolo Rosso, and José Miguel Benedíruiz. 2007. Anersys: An arabic named entity recognition system based on maximum entropy. In *Computational Linguistics and Intelligent Text Processing: 8th International Conference, CICLing 2007, Mexico City, Mexico, February 18-24, 2007. Proceedings 8*, pages 143–153. Springer.

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

Nada Boudjellal, Huaping Zhang, Asif Khan, Arshad Ahmad, Rashid Naseem, Jianyun Shang, and Lin Dai. 2021. Abioner: a bert-based model for arabic biomedical named-entity recognition. *Complexity*, 2021:1–6.

ElMehdi Boujou, Hamza Chataoui, Abdellah El Mekki, Saad Benjelloun, Ikram Chairi, and Ismail Berrada. 2021. An open access nlp dataset for arabic dialects: Data collection, labeling, and model construction. *arXiv preprint arXiv:2102.11000*.

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.

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*.

Omar Einea, Ashraf Elnagar, and Ridhwan Al Debsi. 2019. Sanad: Single-label arabic news articles dataset for automatic text categorization. *Data in brief*, 25:104076.

Mahmoud El-Haj and Rim Koulali. 2013. Kalimat a multipurpose arabic corpus. In *Second workshop on Arabic corpus linguistics (WACL-2)*, pages 22–25.

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*.

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.

Sonja Häffner, Martin Hofer, Maximilian Nagl, and Julian Walterskirchen. 2023. Introducing an interpretable deep learning approach to domain-specific dictionary creation: A use case for conflict prediction. *Political Analysis*, pages 1–19.

Yibo Hu, MohammadSaleh Hosseini, Erick Skorupa Parolin, Javier Osorio, Latifur Khan, Patrick Brandt, and Vito 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*, pages 5469–5482.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

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

Rim Laatar, Chafik Aloulou, and Lamia Hadrich Belghuith. 2018. Word2vec for arabic word sense disambiguation. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23*, pages 308–311. Springer.

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.

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*.

Khouloud Meskaldji, Salim Chikhi, and Imene Bensalem. 2018. A new multi varied arabic corpus. In *2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, pages 1–5. IEEE.

BINIZ mohamed. 2018. Dataset for arabic classification.

Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A Smith. 2012. Recall-oriented learning of named entities in arabic wikipedia. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 162–173.

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.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for ner. *arXiv preprint arXiv:1902.00193*.

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.

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.

David Webber, Arie Kruglanski, Erica Molinario, and Katarzyna Jasko. 2020. Ideologies that justify political violence. *Current Opinion in Behavioral Sciences*, 34:107–111.

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.

Wooseong Yang, Sultan Alsarra, Luay Abdeljaber, Niamat Zawad, Zeinab Delaram, Javier Osorio, Latifur Khan, Patrick Brandt, and Vito D'Orazio. 2023. Conflibert-spanish: A pre-trained spanish language model for political conflict and violence. In *Proceedings of The 5th IEEE Conference on "Machine Learning and Natural Language Processing: Models, Systems, Data and Applications"*.