Computational and Mathematical Organization Theory
https://doi.org/10.1007/s10588-025-09408-2

ORIGINAL PAPER

Multilingual approaches to extractive question answering in political texts

Sultan Alsarra[1] · Mubarak Alrashoud[1] · Javier Osorio[2] · Vito D'Orazio[3] · Latifur Khan[4] · Patrick T. Brandt[5]

Received: 10 December 2024 / Accepted: 26 August 2025

Abstract

This work contributes to multilingual extractive question answering (QA) by presenting QA-specialized versions of ConfliBERT for English, Spanish, and Arabic, language models designed for analyzing political conflict and violence. A cross-lingual QA framework is proposed, including curated datasets and the Spanish translation of an English QA corpus to mitigate the scarcity of annotated resources for low-resource languages. The models are fine-tuned specifically for extractive QA and benchmarked against general-purpose BERT variants, showing consistent gains across all target languages. By addressing language gaps in high-stakes domains, the study underscores the potential of multilingual QA systems to support both research and decision-making in political contexts.

Keywords Extractive QA · Multilingual question answering · Political violence · Spanish NLP · Arabic NLP · Domain-specific models · Natural language understanding · Large language models

1 Introduction

The ConfliBERT family consists of domain-adapted large language models (LLMs) built to process text related to political violence and armed conflict across multiple languages, including English (Hu et al. 2022), Spanish (Yang et al. 2023), and Arabic (Alsarra et al. 2023). Based on specially curated corpora, these models were originally trained for tasks such as binary classification, named entity recognition (NER), multi-class and multi-label classification. This work expands their scope to include question-answering (QA) tasks, with a focus on extractive QA in conflict-related texts. Through this extension, the models address more linguistically complex NLP tasks within specialized domains.

Extended author information available on the last page of the article

Multilingual extractive QA systems encounter unique challenges stemming from structural and syntactic variation across languages, along with uneven dataset availability and domain specificity. Researchers (Lee et al. 2016) distinguish among three core QA types: extractive, open-generative, and closed-generative. While the generative variants synthesize answers from internal model knowledge, extractive QA retrieves answers directly from a given context. In the case of news reporting on conflict in the Middle East, for instance, extractive QA might locate key facts such as participating armed groups, types of tactics being used, or casualty counts. Such capabilities are essential for conflict researchers and analysts who require reliable tools to extract factual information from text.

While generative QA models are increasingly common, their lack of grounding in source text can lead to factual errors in sensitive settings. A motivating example from ChatGPT, shown in Section 5.4.2, illustrates this: the model correctly predicted the date Hosni Mubarak assumed power but incorrectly attributed it to a resignation rather than an assassination. This kind of factual inaccuracy, even in otherwise capable systems, underscores the risks of relying on generative approaches in politically sensitive domains, which reinforces the need for extractive, context-grounded QA.

This study presents QA-specialized versions of ConfliBERT for English, Spanish, and Arabic, fine-tuned for extractive question answering in the domain of political conflict and violence. Using curated datasets and language-specific fine-tuning techniques, these models outperform their baseline BERT counterparts in domain-focused QA. The multilingual design enables robust performance across three languages, helping to address the lack of QA tools in under-resourced linguistic settings.

This work offers three main contributions. First, it applies language-aware fine-tuning to address structural and linguistic differences in English, Spanish, and Arabic for extractive QA. Second, it draws on seven curated QA datasets (two English, two Spanish, and three Arabic), including a Spanish-translated version of NewsQA to mitigate dataset scarcity. Finally, the study evaluates performance on multilingual QA benchmarks, showing the benefits of domain-specific modeling in conflict-focused applications, and demonstrates real-world applicability through multilingual, human-translated United Nations political texts from the UN Parallel Corpus (UNPC).

These contributions aim to assist researchers, practitioners, and analysts working in the field of political conflict monitoring, where access to multilingual and factually grounded information is essential.

2 Background and motivation

The release of BERT (Devlin et al. 2018) marked a turning point in NLP, enabling significant progress in specialized applications and improving outcomes in complex tasks such as question answering (QA). Yet, while advances in domain adaptation and QA have progressed independently, efforts to combine these directions remain limited, particularly in the context of domain-specific QA.

2.1 Domain-specific language models

While foundational to NLP, BERT was originally trained on broad, non-specialized text. This limits its ability to capture terminology and structure in fields that rely on technical language. To overcome this, researchers introduced domain-adapted BERT variants, each pre-trained on domain-specific corpora to improve performance in specialized tasks. Empirical studies show that such models consistently outperform general-purpose BERT in their respective domains (Chalkidis et al. 2020; Rasmy et al. 2021).

For instance, LegalBERT (Chalkidis et al. 2020) specializes in legal text and achieves better performance than baseline BERT across multiple legal benchmarks. Likewise, Med-BERT (Rasmy et al. 2021), trained on millions of clinical records, improves accuracy on health-related tasks. These domain-tuned models highlight the utility of tailored pretraining in specialized applications.

Drawing on this momentum, ConfliBERT (Hu et al. 2022) was introduced as a domain-specific variant focused on political conflict and violence. Unlike traditional rule-based systems (Schrodt 2006, 2009; Osorio and Reyes 2017), which rely on expert-defined heuristics, ConfliBERT leverages large-scale, labeled data to learn political patterns through deep learning. Empirical results show it outperforms alternative models in this domain (Hu et al. 2022; Häffner et al. 2023). Multilingual adaptations of ConfliBERT have also been developed to process conflict-related text in Spanish (Yang et al. 2023) and Arabic (Alsarra et al. 2023).

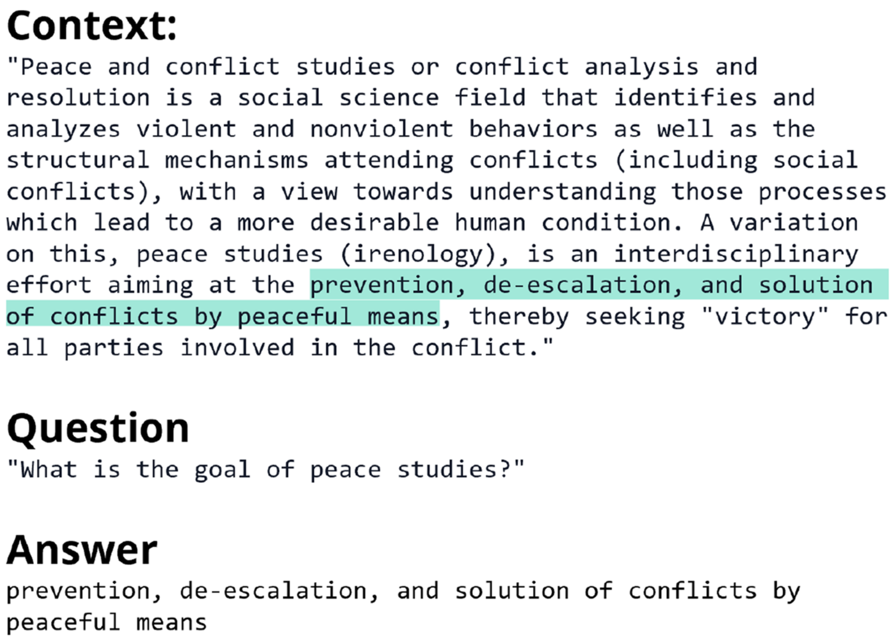
ConfliBERT variants differ based on their pretraining strategy and backbone architecture. Continual pretraining starts from an existing BERT checkpoint (e.g., multilingual BERT Pires et al. 2019, BETO Cañete et al. 2023, AraBERT Antoun et al. 2020) and continues training on a large political conflict corpus, retaining general language knowledge while adapting to the domain. From-scratch pretraining builds a new vocabulary and trains entirely on domain text, which can yield stronger specialization when large unlabeled data is available.

For Spanish, both multilingual BERT and BETO backbones (cased and uncased) are used; for Arabic, multilingual BERT and AraBERT are used. This diversity enables direct comparison between language-specific and multilingual architectures, as well as between continual and from-scratch approaches. Despite these developments, no ConfliBERT variant—whether English, Spanish, or Arabic—had been fine-tuned or systematically evaluated for extractive question answering, a gap this study directly addresses.

2.2 Limitations of extractive QA

Extractive question answering is considered a challenging downstream task for transformer-based models (Rajpurkar et al. 2016). In this setting, the model receives both a question and a passage of context, and must identify a span of text that directly answers the question. Figure 1 illustrates this setup.

Fine-tuning for extractive QA involves feeding the question and the corresponding context into the model, separated by a [SEP] token. The model then predicts the start and end positions of the answer span within the input text. During training, BERT



Context:

"Peace and conflict studies or conflict analysis and resolution is a social science field that identifies and analyzes violent and nonviolent behaviors as well as the structural mechanisms attending conflicts (including social conflicts), with a view towards understanding those processes which lead to a more desirable human condition. A variation on this, peace studies (irenology), is an interdisciplinary effort aiming at the prevention, de-escalation, and solution of conflicts by peaceful means, thereby seeking "victory" for all parties involved in the conflict."

Question

"What is the goal of peace studies?"

Answer

prevention, de-escalation, and solution of conflicts by peaceful means

Fig. 1 Extractive QA Example

learns to associate specific tokens with accurate answers based on supervision from the labeled data.

A key obstacle in developing domain-specific extractive QA systems is the scarcity of annotated datasets, especially for low-resource languages like Spanish and Arabic. To address this, seven datasets were curated or adapted for this study: two in English, two in Spanish, and three in Arabic. These resources required careful preprocessing to ensure format consistency and relevance to the conflict domain. Each dataset provided QA pairs for model fine-tuning and evaluation. Performance was measured against general-purpose BERT baselines, enabling comparison between domain-specific and standard models.

3 Datasets and QA pipeline preparation

To support extractive QA, the study prepared datasets tailored to the linguistic properties of English, Spanish, and Arabic. The goal was to ensure alignment with the ConfliBERT models and their intended use in political conflict research. The datasets were selected or curated to reflect this domain, enabling evaluation in realistic multilingual applications.

3.1 Preprocessing and fine-tuning infrastructure

This study is based on custom preprocessing scripts developed in Python using the HuggingFace Transformers framework. These scripts enabled automated fine-tuning and evaluation pipelines for multilingual QA. To adapt to linguistic variation, the study implemented language-specific preprocessing steps: English and Spanish texts used lowercase inputs and were stripped of punctuation, articles, and extra whitespace, while Arabic text underwent additional normalization such as diacritic (tashkeel) and elongation removal, which are not critical for answer span extraction.

Initially in TSV format, the methodology converted the datasets into HuggingFace-compatible JSON files using a custom parser designed to standardize field names and ensure format consistency across all languages.

The fine-tuning relied on parallel computing using multi-GPU setups, with each job reading configuration details from structured JSON argument files. The infrastructure included four A100 GPUs, 64 CPUs, and 248 GB of RAM, which supported large-scale experimentation and reduced training time.

The training and validation data splits use a 60–40 ratio for all datasets except ARCD, which followed its official 50–50 split (702 QA pairs for testing, remainder for training). The format followed a standardized structure with five fields: ID (a unique identifier), Title (document category), Question, Context, and Answers. The Answers field included two keys: "Answer_Start" for token start indices, and "Text" for the answer span. This format ensured compatibility with HuggingFace's QA pipelines and allowed for consistent multilingual fine-tuning.

To ensure answer span validity, the methodology filtered out QA pairs where the labeled answer did not align with any contiguous substring in the context, or where the question lacked a clearly identifiable answer. This filtering process resulted in the exclusion of approximately 15–25% of QA pairs across all datasets. For example, the Arabic datasets went down from around 3,800 to 2,950 items, and the Spanish sets from approximately 2,600 to 2,300 items.

3.2 English QA datasets

The study uses two QA datasets for English: NewsQA (Trischler et al. 2017) and SQuAD v1.1 (Rajpurkar et al. 2016). NewsQA is an extractive QA dataset with over 120,000 question-answer pairs drawn from CNN news articles, including significant coverage of political conflict and violence. SQuAD v1.1 is a benchmark dataset with more than 100,000 question-answer pairs from Wikipedia. Although not domain-specific, it is widely used to evaluate QA models due to its structure, clarity, and linguistic variety.

SQuAD v1.1 was selected instead of v2.0 to maintain focus on high-quality, answerable QA pairs. Although the dataset was already well-structured, domain-specific filtering and formatting adjustments were applied to align it with multilingual QA pipelines. To filter for domain relevance, QA pairs associated with paragraphs containing keywords such as "war," "military," "protest," and "uprising" were retained, resulting in a conflict-specific subset of approximately 2,300 examples. This subset was converted into HuggingFace's DatasetDict format to ensure consistency across English, Spanish, and Arabic datasets.

NewsQA required more extensive preparation due to formatting inconsistencies and crowd-sourced variation. Using the retrieval script provided with the dataset, article IDs were mapped to the corresponding CNN texts. Low-quality QA pairs, those where the answer did not appear in the context or where alignment was ambiguous, were removed based on original quality flags. This filtering reduced the dataset from over 120,000 to approximately 25,000 well-formed QA pairs, improving dataset consistency and reliability for model evaluation. Both SQuAD v1.1 and NewsQA are publicly available datasets.

3.3 Spanish QA datasets

This study uses two QA datasets for Spanish: a translated version of NewsQA and the native Spanish SQAC dataset (Gutiérrez-Fandiño et al. 2021). Since NewsQA exists only in English, we translated it into Spanish using the Translate Align Retrieve (TAR) (Carrino et al. 2019) method, which has proven effective in adapting extractive QA datasets such as SQuAD (Rajpurkar et al. 2016). The TAR framework was selected for its capacity to maintain semantic alignment between source and target texts, yielding high-quality translations suitable for QA applications.

Preprocessing and Cleaning The NewsQA dataset was downloaded using official scripts and prepared for QA formatting. We removed duplicated or low-quality QA pairs and structured the output into five fields: ID, title, context, question, and answers. The formatted version was saved using HuggingFace's DatasetDict format for consistency across all languages and integration with the fine-tuning pipeline.

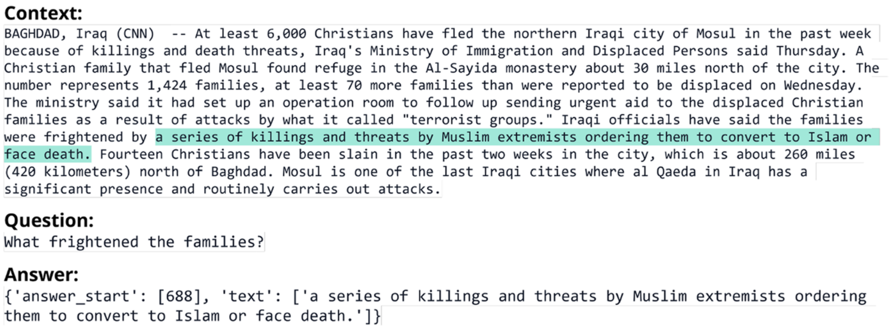
Translation The TAR pipeline includes three stages: machine translation, word alignment, and answer projection. To avoid inaccuracies from long CNN contexts, each article was split into sentences using the NLTK toolkit (Bird and Loper 2004). These were translated into Spanish using the opus-mt-en-esneural model (Tiedemann et al. 2023). Questions were translated directly without tokenization to preserve answer span fidelity.

Next, SimAlign (Jalili Sabet et al. 2020) was used to align tokens across language pairs using multilingual BERT embeddings, with the ArgMax mode for high-precision matches. Sentence-level alignment helped preserve structure, and multi-sentence answers were merged and realigned to ensure positional accuracy.

Finally, text retrieval inserted the aligned Spanish answer spans into the translated context. Index calculations were used to resolve alignment gaps and compute the correct answer start locations. The resulting dataset was stored using the HuggingFace DatasetDict format. Examples are shown in Figs. 2 and 3. Note that due to copyright restrictions on the original English NewsQA dataset, the translated Spanish version used in this study cannot be released publicly.

The Spanish Question Answering Corpus (SQAC) (Gutiérrez-Fandiño et al. 2021) is a native-language extractive QA dataset for Spanish, containing 6,247 contexts and 18,817 questions, each annotated with one to five answer spans. All texts were originally written in Spanish and sourced from Spanish Wikipedia, Wikinews, and the AnCora corpus, offering a mix of encyclopedic, journalistic, and literary content.

Annotation guidelines were adapted from SQuAD v1.1, and all annotators were native Spanish speakers with linguistic training. The dataset excludes unanswerable



Context:

```
BAGHDAD, Iraq (CNN)  -- At least 6,000 Christians have fled the northern Iraqi city of Mosul in the past week
because of killings and death threats, Iraq's Ministry of Immigration and Displaced Persons said Thursday. A
Christian family that fled Mosul found refuge in the Al-Sayida monastery about 30 miles north of the city. The
number represents 1,424 families, at least 70 more families than were reported to be displaced on Wednesday.
The ministry said it had set up an operation room to follow up sending urgent aid to the displaced Christian
families as a result of attacks by what it called "terrorist groups." Iraqi officials have said the families
were frightened by a series of killings and threats by Muslim extremists ordering them to convert to Islam or
face death. Fourteen Christians have been slain in the past two weeks in the city, which is about 260 miles
(420 kilometers) north of Baghdad. Mosul is one of the last Iraqi cities where al Qaeda in Iraq has a
significant presence and routinely carries out attacks.
```

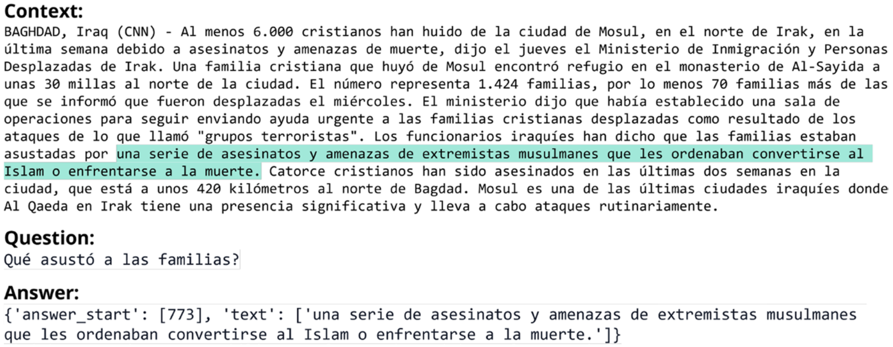
Question:

```
What frightened the families?
```

Answer:

```
{'answer_start': [688], 'text': ['a series of killings and threats by Muslim extremists ordering
them to convert to Islam or face death.']}
```

Fig. 2 NewsQA example (English version)



Context:

```
BAGHDAD, Iraq (CNN) - Al menos 6.000 cristianos han huido de la ciudad de Mosul, en el norte de Irak, en la
última semana debido a asesinatos y amenazas de muerte, dijo el jueves el Ministerio de Inmigración y Personas
Desplazadas de Irak. Una familia cristiana que huyó de Mosul encontró refugio en el monasterio de Al-Sayida a
unas 30 millas al norte de la ciudad. El número representa 1.424 familias, por lo menos 70 familias más de las
que se informó que fueron desplazadas el miércoles. El ministerio dijo que había establecido una sala de
operaciones para seguir enviando ayuda urgente a las familias cristianas desplazadas como resultado de los
ataques de lo que llamó "grupos terroristas". Los funcionarios iraquíes han dicho que las familias estaban
asustadas por una serie de asesinatos y amenazas de extremistas musulmanes que les ordenaban convertirse al
Islam o enfrentarse a la muerte. Catorce cristianos han sido asesinados en las últimas dos semanas en la
ciudad, que está a unos 420 kilómetros al norte de Bagdad. Mosul es una de las últimas ciudades iraquíes donde
Al Qaeda en Irak tiene una presencia significativa y lleva a cabo ataques rutinariamente.
```

Question:

```
Qué asustó a las familias?
```

Answer:

```
{'answer_start': [773], 'text': ['una serie de asesinatos y amenazas de extremistas musulmanes
que les ordenaban convertirse al Islam o enfrentarse a la muerte.']}
```

Fig. 3 NewsQA example (translated to Spanish)

questions and includes a substantial portion of political and government-related QA pairs, making it well-suited for this study's focus. No filtering was applied; the full dataset was used in its original form. It was also converted to the HuggingFace DatasetDict format for multilingual pipeline compatibility. The SQAC dataset is publicly available through HuggingFace (Gutiérrez-Fandiño et al. 2021).

3.4 Arabic QA datasets

For Arabic, the study uses three publicly available QA datasets: XQuAD (Artetxe et al. 2019), MLQA (Lewis et al. 2019), and ARCD (Mozannar et al. 2019). Each provides extractive QA examples relevant to political or governmental topics and supports evaluation across different linguistic sources.

The XQuAD (Cross-lingual Question Answering Dataset) was developed by Google DeepMind to evaluate cross-lingual QA performance. It contains 240 paragraphs and 1,190 QA pairs translated from English to multiple languages, including Arabic. For this study, the Arabic split prepared by the XTREME benchmark (Hu et al. 2020) was used. Since the original source is SQuAD v1.1 (Rajpurkar et al. 2016), the questions retain structural consistency and include a sizable proportion related to political content.

The MLQA (Multilingual Question Answering) dataset, released by Facebook AI, includes over 5,000 QA pairs in seven languages, including Arabic. It follows the SQuAD format but sources context passages directly from native-language Wikipedia articles, many of which discuss political institutions, historical events, and governmental figures. Its multilingual structure makes it particularly useful for cross-lingual QA benchmarking.

The Arabic Reading Comprehension Dataset (ARCD) consists of 1,395 QA pairs collected from Arabic Wikipedia. The dataset covers a range of answer types, including numerical, noun phrase, and descriptive answers, and includes metadata for linguistic complexity, ambiguity, and multi-sentence reasoning. For evaluation, a subset of 702 questions across 78 articles was selected as the test set. Many of the articles focus on political entities and events, aligning the dataset with the objectives of domain-specific QA evaluation.

4 Experimental setup

In total, the empirical foundation of this study relies on 40 models fine-tuned across seven datasets: 12 for English, 16 for Spanish, and 12 for Arabic. For English, the ConfliBERT model types include Continual-cased, Continual-uncased, Scratch-cased, and Scratch-uncased, each evaluated alongside corresponding BERT baselines over two datasets. For Spanish, the ConfliBERT configurations include multilingual-cased, multilingual-uncased, BETO-cased, and BETO-uncased, each compared to its respective BERT-based benchmark across two datasets. For Arabic, the fine-tuned models include versions based on AraBERT and multilingual-uncased backbones, along with their associated baselines, evaluated over three datasets.

To assess performance, the evaluation uses two established extractive QA metrics: Exact Match (EM) and F1 Score. EM measures the percentage of predictions whose start and end tokens exactly match the annotated gold span (i.e., the ground-truth answer span labeled by humans). It is defined as follows:

$$\mathrm{EM} = \frac{1}{N}\sum_{i=1}^{N} 1\left(\hat{a}_i = a_i\right)$$

where N is the number of examples, $\hat{a}_i$ is the model prediction, and a_i is the gold answer span. This highly restrictive metric captures whether a model can align precisely with the dataset's expected answer style.

In addition, the F1 Score offers a more flexible view by measuring token-level overlap between prediction and reference. It is defined as:

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

where precision is the proportion of predicted tokens appearing in the gold answer, and recall is the proportion of gold tokens correctly predicted. F1 is particularly use-

Table 1 English Results

Model Name	Variant	(a) Extractive QA		(b) NewsQA		(c) SQuAD	
		F1 Score	EM Score	F1 Score	EM Score	F1 Score	EM Score
ConfliBERT English	Cont-Cased	80.36	65.76	71.20	49.03	**89.52**	**82.48**
	Cont-Uncased	80.10	65.42	71.56	49.78	88.63	81.05
	Scr-Cased	**80.64**	**65.97**	**72.90**	**50.65**	88.38	81.29
	Scr-Uncased	80.01	65.37	71.22	49.05	88.79	81.68
BERT	Base-Cased	78.63	63.19	68.91	45.67	88.34	80.70
	Base-Uncased	78.53	63.44	68.90	46.39	88.15	80.48

Note: Bold font indicates top performing model

Table 2 Spanish Results

Model Name	Variant	(a) Extractive QA		(b) News QA		(c) SQAC	
		F1 Score	EM Score	F1 Score	EM Score	F1 Score	EM Score
ConfliBERT-Spanish	Cased	70.14	48.00	62.76	33.04	77.51	62.88
	Uncased	69.92	47.90	63.01	33.38	76.83	62.39
	BETO-Cased	**72.30**	**50.21**	64.88	35.08	**79.72**	**65.34**
	BETO-Uncased	72.15	50.16	**65.53**	**35.19**	78.77	65.12
BERT	Cased	69.85	44.16	59.74	30.70	72.96	57.62
	Uncased	66.61	43.98	60.19	30.06	73.02	57.89
	BETO-Cased	71.20	48.85	63.39	33.64	79.00	64.06
	BETO-Uncased	65.71	43.78	59.60	30.47	71.82	57.08

Note: Bold font indicates top performing model

ful for assessing partial matches, especially for longer or multi-token entities common in political and conflict-focused texts, such as “Supreme Council of the Armed Forces.”

The motivation behind selecting EM and F1 is that they are widely adopted and standardized for extractive QA tasks, facilitating direct comparison with existing baselines. While EM captures strict correctness, F1 conveys how much relevant content is retained in a prediction.

All models were trained using identical hyper-parameters to ensure consistency across experiments. Following best practices for BERT-based fine-tuning (Devlin et al. 2018), each model was trained for five epochs using five random seeds. The training setup used a batch size of 8, learning rate of 5e-5, maximum answer length of 100 tokens, and a context window capped at 384 tokens.

5 Results and analysis

The results in Tables 1, 2, and 3 demonstrate that domain-specialized ConfliBERT models consistently outperform general-purpose BERT baselines in English, Spanish, and Arabic QA datasets. While the magnitude of improvement varies by dataset and language, gains are observed across all evaluation settings. These improvements are especially meaningful in the context of extractive QA on political texts, where

Table 3 Arabic Results

Model Name	Variant	(a) Extractive QA		(b) MLQA		(c) XQUAD		(d) ARCD	
		F1 Score	EM Score	F1 Score	EM Score	F1 Score	EM Score	F1 Score	EM Score
ConfliBERT-Arabic	AraBERT	**61.90**	**40.11**	**64.86**	**44.24**	**63.33**	**47.19**	**57.43**	**28.92**
	Uncased	60.76	37.79	64.11	43.47	62.21	46.10	55.95	23.79
BERT	AraBERT	60.18	38.64	63.41	42.95	62.29	46.20	54.84	26.78
	Uncased	58.35	35.50	62.16	41.00	60.55	44.54	52.33	20.94

Note: Bold font indicates top performing model

even modest increases in Exact Match (EM) or F1 Score can significantly enhance answer quality. Each subsection below presents a focused analysis of results for English, Spanish, and Arabic.

5.1 ConfliBERT-english QA

Table 1 presents the performance of English models across three datasets. ConfliBERT-English models consistently outperformed both cased and uncased BERT baselines in F1 Score and Exact Match. The Scr-Cased variant delivered the strongest results on NewsQA, achieving 72.90 (F1) and 50.65 (EM). For SQuAD, the Cont-Cased model led with an F1 of 89.52 and EM of 82.48, underscoring ConfliBERT's robustness on well-structured English QA tasks. These findings support the value of domain-specific fine-tuning for conflict-focused extractive QA in English.

5.2 ConfliBERT-spanish QA

Table 2 presents results for Spanish-language models fine-tuned for extractive QA. ConfliBERT-Spanish variants consistently outperformed corresponding BERT-based baselines across all datasets. The BETO-Cased variant obtained the highest average F1 (72.30) and Exact Match (50.21), reflecting the strength of domain-specific fine-tuning on a Spanish-native model. On NewsQA, the BETO-Uncased model achieved the best F1 (65.53) and EM (35.19). On SQAC, BETO-Cased again performed best, reaching 79.72 F1 and 65.34 EM, outperforming all other variants and confirming its effectiveness for Spanish QA tasks.

5.3 ConfliBERT-arabic QA

Table 3 reports performance for Arabic-language QA models. ConfliBERT-Arabic variants consistently outperformed their respective BERT-based baselines across all datasets. The AraBERT variant yielded the strongest overall performance, with an average F1 of 61.90 and Exact Match of 40.11, highlighting the gains of domain-specific fine-tuning. On MLQA, it achieved 64.86 F1 and 44.24 EM, outperforming all other variants. Performance remained strongest on XQUAD and ARCD as well, with 63.33 and 57.43 F1, respectively, confirming its effectiveness for conflict-focused extractive QA in Arabic.

5.4 Qualitative evaluation of answers

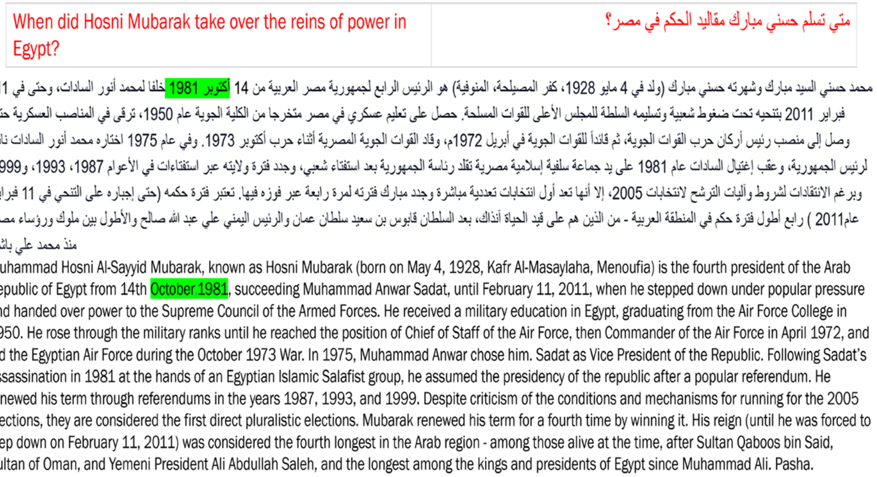
To supplement the quantitative results, this section presents illustrative examples comparing the output of ConfliBERT-Arabic and its BERT-based baseline. A brief comparison with ChatGPT (Ray 2023) is also included to highlight how different models interpret political context and select answers. These examples illustrate model behavior and are not intended as formal evaluation. For clarity, examples originally in Arabic were translated into English by native speakers to aid interpretation, while all evaluations were conducted on the original Arabic texts.

5.4.1 ConfliBERT vs BERT baseline

The first example focuses on a question related to former Egyptian president Hosni Mubarak. Question Q1, asked in Arabic, was: “When did Hosni Mubarak take over the reins of power in Egypt?” The input context, shown in Fig. 4, includes the annotated gold answer span “October 1981,” highlighted in both Arabic and its English translation.

ConfliBERT-Arabic predicted “1981,” correctly identifying the core information while omitting the month. In contrast, the BERT baseline returned “1950,” which refers to Mubarak’s Air Force graduation, not his assumption of power. These predictions are shown in Fig. 5.

A follow-up question, Q2, asked: “To whom did Hosni Mubarak hand power after the 2011 protests?” The correct answer was “to the Supreme Council of the Armed Forces.” Figure 6 shows the supporting context and annotated gold answer. ConfliBERT-Arabic predicted the correct span, while the BERT baseline returned “11 February 2022,” which incorrectly refers to the date of Mubarak’s resignation rather than the transfer of power. These predictions are shown in Fig. 7.

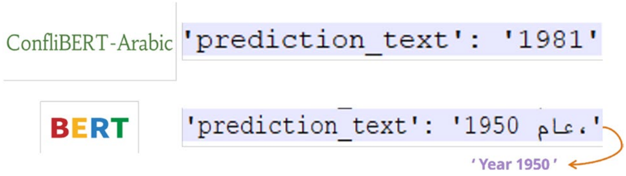


When did Hosni Mubarak take over the reins of power in Egypt?	متي تسلم حسني مبارك مقاليد الحكم في مصر؟

محمد حسني السيد مبارك وشهرته حسني مبارك (ولد في 4 مايو 1928، كفر المصيلحة، المنوفية) هو الرئيس الرابع لجمهورية مصر العربية من 14 أكتوبر 1981 خلفا لمحمد أنور السادات، وحتى في 11 فبراير 2011 بتنحيه تحت ضغوط شعبية وتسليمه السلطة للمجلس الأعلى للقوات المسلحة. حصل على تعليم عسكري في مصر متخرجا من الكلية الجوية عام 1950، ترقى في المناصب العسكرية حتى وصل إلى منصب رئيس أركان حرب القوات الجوية، ثم قائداً للقوات الجوية في أبريل 1972م، وقاد القوات الجوية المصرية أثناء حرب أكتوبر 1973. وفي عام 1975 اختاره محمد أنور السادات نائباً لرئيس الجمهورية، وعقب إغتيال السادات عام 1981 على يد جماعة سلفية إسلامية مصرية تقلد رئاسة الجمهورية بعد استفتاء شعبي، وجدد فترة ولايته عبر استفتاءات في الأعوام 1987، 1993، و1999 وبرغم الانتقادات لشروط وآليات الترشح لانتخابات 2005، إلا أنها تعد أول انتخابات تعددية مباشرة وجدد مبارك فترته لمرة رابعة عبر فوزه فيها. تعتبر فترة حكمه (حتى إجباره على التنحي في 11 فبراير عام2011) رابع أطول فترة حكم في المنطقة العربية - من الذين هم على قيد الحياة آنذاك، بعد السلطان قابوس بن سعيد سلطان عمان والرئيس اليمني علي عبد الله صالح والأطول بين ملوك ورؤساء مصر منذ محمد علي باشا.

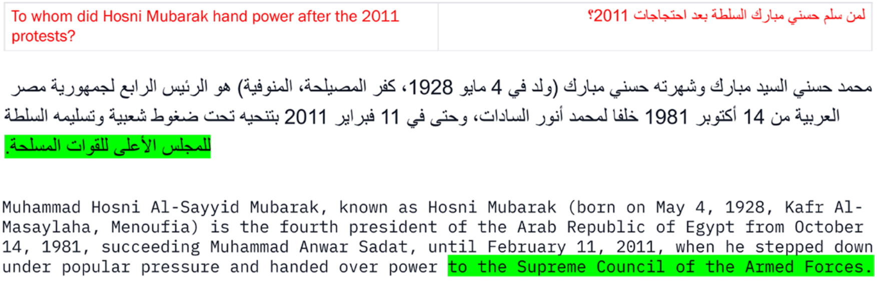
Muhammad Hosni Al-Sayyid Mubarak, known as Hosni Mubarak (born on May 4, 1928, Kafr Al-Masaylaha, Menoufia) is the fourth president of the Arab Republic of Egypt from 14th October 1981, succeeding Muhammad Anwar Sadat, until February 11, 2011, when he stepped down under popular pressure and handed over power to the Supreme Council of the Armed Forces. He received a military education in Egypt, graduating from the Air Force College in 1950. He rose through the military ranks until he reached the position of Chief of Staff of the Air Force, then Commander of the Air Force in April 1972, and led the Egyptian Air Force during the October 1973 War. In 1975, Muhammad Anwar chose him. Sadat as Vice President of the Republic. Following Sadat’s assassination in 1981 at the hands of an Egyptian Islamic Salafist group, he assumed the presidency of the republic after a popular referendum. He renewed his term through referendums in the years 1987, 1993, and 1999. Despite criticism of the conditions and mechanisms for running for the 2005 elections, they are considered the first direct pluralistic elections. Mubarak renewed his term for a fourth time by winning it. His reign (until he was forced to step down on February 11, 2011) was considered the fourth longest in the Arab region - among those alive at the time, after Sultan Qaboos bin Said, Sultan of Oman, and Yemeni President Ali Abdullah Saleh, and the longest among the kings and presidents of Egypt since Muhammad Ali. Pasha.

Fig. 4 Q1: Context, Question, and Highlighted Answer



```
'prediction_text': '1981'
'prediction_text': '1950 عام،'
```

Fig. 5 Q1: Predicted Answers from ConfliBERT-Arabic and Base BERT



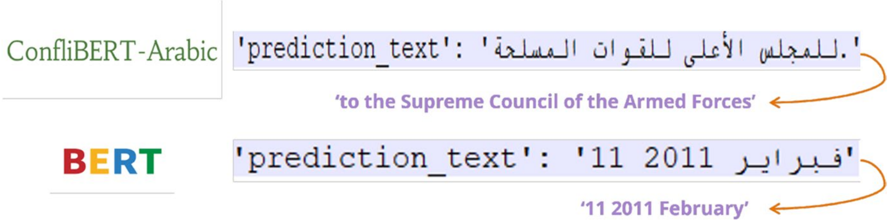
To whom did Hosni Mubarak hand power after the 2011 protests?

لمن سلم حسني مبارك السلطة بعد احتجاجات 2011؟

محمد حسني السيد مبارك وشهرته حسني مبارك (ولد في 4 مايو 1928، كفر المصيلحة، المنوفية) هو الرئيس الرابع لجمهورية مصر العربية من 14 أكتوبر 1981 خلفا لمحمد أنور السادات، وحتى في 11 فبراير 2011 بتنحيه تحت ضغوط شعبية وتسليمه السلطة للمجلس الأعلى للقوات المسلحة.

```
Muhammad Hosni Al-Sayyid Mubarak, known as Hosni Mubarak (born on May 4, 1928, Kafr Al-
Masaylaha, Menoufia) is the fourth president of the Arab Republic of Egypt from October
14, 1981, succeeding Muhammad Anwar Sadat, until February 11, 2011, when he stepped down
under popular pressure and handed over power to the Supreme Council of the Armed Forces.
```

Fig. 6 Q2: Context, Question and Highlighted Answer



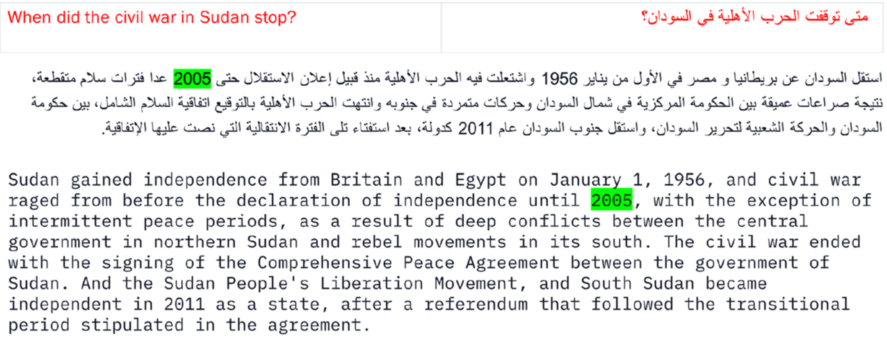
```
'prediction_text': 'للمجلس الأعلى للقوات المسلحة.'
'prediction_text': '11 2011 فبراير'
```

Fig. 7 Q2: Predicted Answers from ConfliBERT-Arabic and Base BERT

The third question, Q3, asked: "When did the civil war in Sudan stop?" The correct answer was "2005," shown in Fig. 8 with supporting context and annotated gold span. ConfliBERT-Arabic correctly identified the answer and included additional context: "from before the declaration of independence until 2005." The BERT baseline returned "1 January 1956," which incorrectly refers to Sudan's independence rather than the end of the civil war. These predictions are shown in Fig. 9.

To evaluate model versatility beyond local or regional topics, a fourth example, Q4, asks: "When did Donald Trump take office as President of the United States?" The correct answer is "January 20, 2017," shown in Fig. 10. ConfliBERT-Arabic correctly identified the target date, while the BERT baseline incorrectly returned "14 June 1946," which corresponds to Trump's birth date rather than his inauguration. Predicted answers for Q4 are shown in Fig. 11.

Overall, the ConfliBERT-Arabic extractive QA models consistently outperformed the baseline BERT models, providing accurate responses across a range of politically focused questions. The BERT baseline struggled in several areas, particularly with interpreting political context and correctly identifying dates and numerical informa-



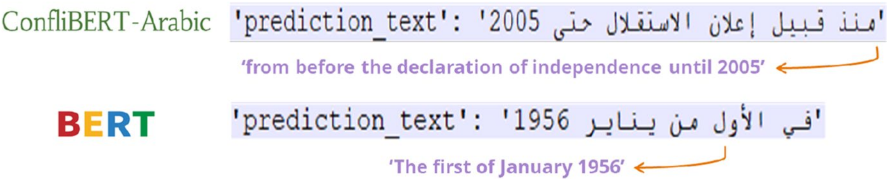
When did the civil war in Sudan stop?

متى توقفت الحرب الأهلية في السودان؟

استقل السودان عن بريطانيا و مصر في الأول من يناير 1956 واشتعلت فيه الحرب الأهلية منذ قبيل إعلان الاستقلال حتى 2005 عدا فترات سلام متقطعة، نتيجة صراعات عميقة بين الحكومة المركزية في شمال السودان وحركات متمردة في جنوبه وانتهت الحرب الأهلية بالتوقيع اتفاقية السلام الشامل، بين حكومة السودان والحركة الشعبية لتحرير السودان، واستقل جنوب السودان عام 2011 كدولة، بعد استفتاء تلى الفترة الانتقالية التي نصت عليها الإتفاقية.

Sudan gained independence from Britain and Egypt on January 1, 1956, and civil war raged from before the declaration of independence until 2005, with the exception of intermittent peace periods, as a result of deep conflicts between the central government in northern Sudan and rebel movements in its south. The civil war ended with the signing of the Comprehensive Peace Agreement between the government of Sudan. And the Sudan People's Liberation Movement, and South Sudan became independent in 2011 as a state, after a referendum that followed the transitional period stipulated in the agreement.

Fig. 8 Q3: Context, Question, and Highlighted Answer



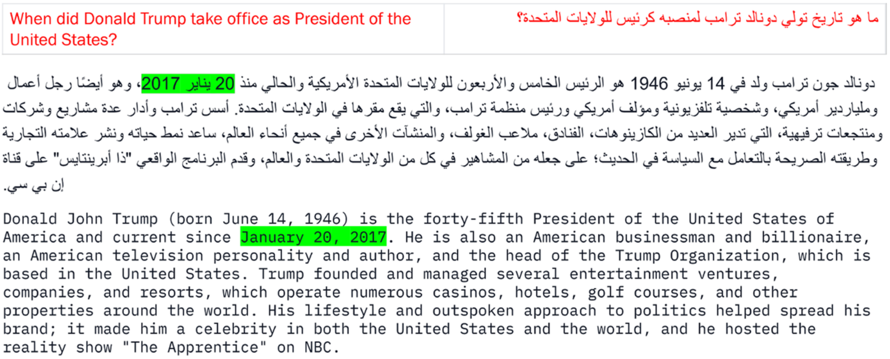
ConfliBERT-Arabic 'prediction_text': 'منذ قبيل إعلان الاستقلال حتى 2005'

'from before the declaration of independence until 2005'

BERT 'prediction_text': 'في الأول من يناير 1956'

'The first of January 1956'

Fig. 9 Q3: Predicted Answers from ConfliBERT-Arabic and Base BERT



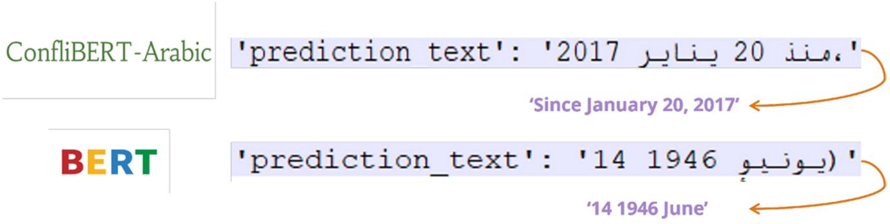
When did Donald Trump take office as President of the United States?

ما هو تاريخ تولي دونالد ترامب لمنصبه كرئيس للولايات المتحدة؟

دونالد جون ترامب ولد في 14 يونيو 1946 هو الرئيس الخامس والأربعون للولايات المتحدة الأمريكية والحالي منذ 20 يناير 2017، وهو أيضًا رجل أعمال وملياردير أمريكي، وشخصية تلفزيونية ومؤلف أمريكي ورئيس منظمة ترامب، والتي يقع مقرها في الولايات المتحدة. أسس ترامب وأدار عدة مشاريع وشركات ومنتجعات ترفيهية، التي تدير العديد من الكازينوهات، الفنادق، ملاعب الغولف، والمنشآت الأخرى في جميع أنحاء العالم، ساعد نمط حياته ونشر علامته التجارية وطريقته الصريحة بالتعامل مع السياسة في الحديث؛ على جعله من المشاهير في كل من الولايات المتحدة والعالم، وقدم البرنامج الواقعي "ذا أبرينتايس" على قناة إن بي سي.

Donald John Trump (born June 14, 1946) is the forty-fifth President of the United States of America and current since January 20, 2017. He is also an American businessman and billionaire, an American television personality and author, and the head of the Trump Organization, which is based in the United States. Trump founded and managed several entertainment ventures, companies, and resorts, which operate numerous casinos, hotels, golf courses, and other properties around the world. His lifestyle and outspoken approach to politics helped spread his brand; it made him a celebrity in both the United States and the world, and he hosted the reality show "The Apprentice" on NBC.

Fig. 10 Q4: Context, Question, and Highlighted Answer



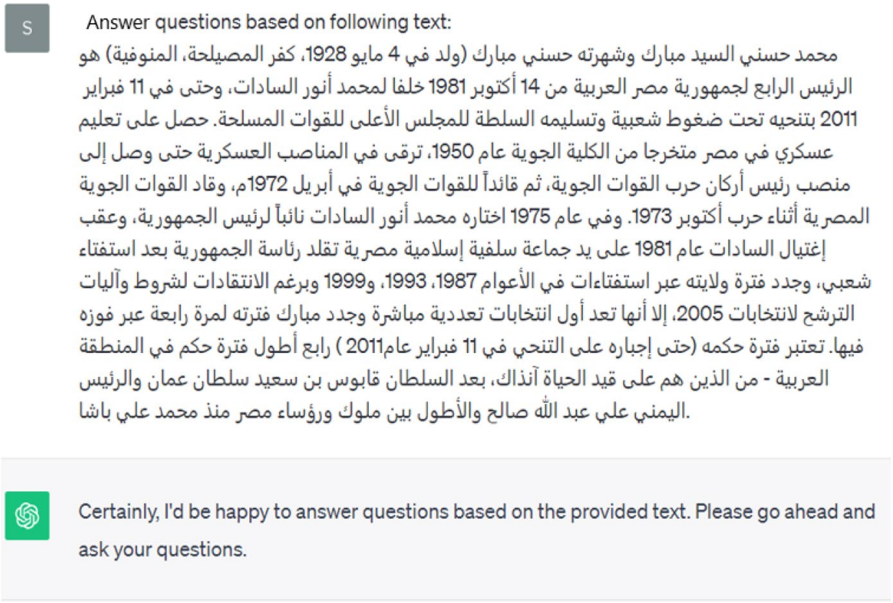
ConfliBERT-Arabic 'prediction text': 'منذ 20 يناير 2017،'

'Since January 20, 2017'

BERT 'prediction_text': '14 1946 يونيو)'

'14 1946 June'

Fig. 11 Q4: Predicted Answers from ConfliBERT-Arabic and Base BERT

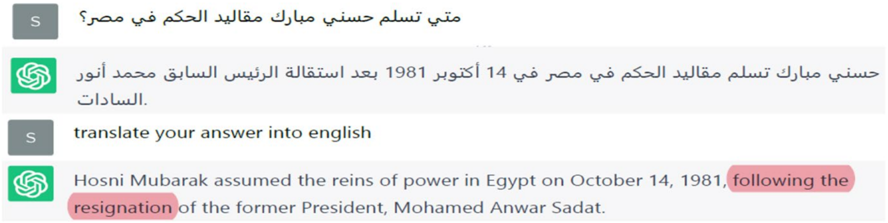


Answer questions based on following text:

محمد حسني السيد مبارك وشهرته حسني مبارك (ولد في 4 مايو 1928، كفر المصيلحة، المنوفية) هو
الرئيس الرابع لجمهورية مصر العربية من 14 أكتوبر 1981 خلفا لمحمد أنور السادات، وحتى في 11 فبراير
2011 بتنحيه تحت ضغوط شعبية وتسليمه السلطة للمجلس الأعلى للقوات المسلحة. حصل على تعليم
عسكري في مصر متخرجا من الكلية الجوية عام 1950، ترقى في المناصب العسكرية حتى وصل إلى
منصب رئيس أركان حرب القوات الجوية، ثم قائداً للقوات الجوية في أبريل 1972م، وقاد القوات الجوية
المصرية أثناء حرب أكتوبر 1973. وفي عام 1975 اختاره محمد أنور السادات نائباً لرئيس الجمهورية، وعقب
إغتيال السادات عام 1981 على يد جماعة سلفية إسلامية مصرية تقلد رئاسة الجمهورية بعد استفتاء
شعبي، وجدد فترة ولايته عبر استفتاءات في الأعوام 1987، 1993، و1999 وبرغم الانتقادات لشروط وآليات
الترشح لانتخابات 2005، إلا أنها تعد أول انتخابات تعددية مباشرة وجدد مبارك فترته لمرة رابعة عبر فوزه
فيها. تعتبر فترة حكمه (حتى إجباره على التنحي في 11 فبراير عام2011) رابع أطول فترة حكم في المنطقة
العربية - من الذين هم على قيد الحياة آنذاك، بعد السلطان قابوس بن سعيد سلطان عمان والرئيس
اليمني علي عبد الله صالح والأطول بين ملوك ورؤساء مصر منذ محمد علي باشا.

Certainly, I'd be happy to answer questions based on the provided text. Please go ahead and ask your questions.

Fig. 12 ChatGPT: Prompt and Q1 Context



متي تسلم حسني مبارك مقاليد الحكم في مصر؟

حسني مبارك تسلم مقاليد الحكم في مصر في 14 أكتوبر 1981 بعد استقالة الرئيس السابق محمد أنور
السادات.

translate your answer into english

Hosni Mubarak assumed the reins of power in Egypt on October 14, 1981, following the resignation of the former President, Mohamed Anwar Sadat.

Fig. 13 ChatGPT: Answer for Q1

tion. In contrast, ConfliBERT-Arabic demonstrated stronger contextual understanding and more reliable handling of temporal and quantitative details.

5.4.2 Comparison with ChatGPT

ChatGPT was evaluated by providing it with the same Q1 context in Arabic and asking: “When did Hosni Mubarak take over the reins of power in Egypt?” ChatGPT returned the answer “October 14, 1981,” partially matching the gold span. However, it inaccurately stated that Mubarak assumed power following the *resignation* of President Anwar Sadat, whereas, in fact, Sadat was assassinated. Figures 12 and 13 show the prompt and ChatGPT’s full response.

ChatGPT occasionally inferred extra details not present in the provided context, leading to factual inaccuracies. In contrast, the ConfliBERT-Arabic model extracted answers directly from the passage, avoiding speculative or hallucinated content.

Given the closed nature of ChatGPT, it is not possible to assess the source of this factual inaccuracy. However, this type of error is especially important for applications in political science and conflict research, where factual precision and source-grounded responses are critical. In political science, there are enormous differences between contexts where the transfer of power occurs peacefully (as in a voluntary resignation) or in a violent manner (as in a political assassination).

6 Multilingual QA evaluation with UNPC

The United Nations Parallel Corpus (UNPC) (Ziemski et al. 2016) was utilized for this study. It is a comprehensive collection of official United Nations (UN) documents spanning 1990–2014, comprising 86,307 documents translated by professional linguists from the UN Department for General Assembly and Conference Management (DGACM) into the six official UN languages: English, Spanish, Arabic, French, Russian, and Chinese. These translations are aligned at the sentence level, resulting in 11,365,709 parallel sentences. The UNPC has been previously used to evaluate domain-specific large language models (LLMs) across native and machine-translated texts in the context of political conflict (Osorio et al. 2024, 2025; Ziemski et al. 2016; Wu et al. 2021).

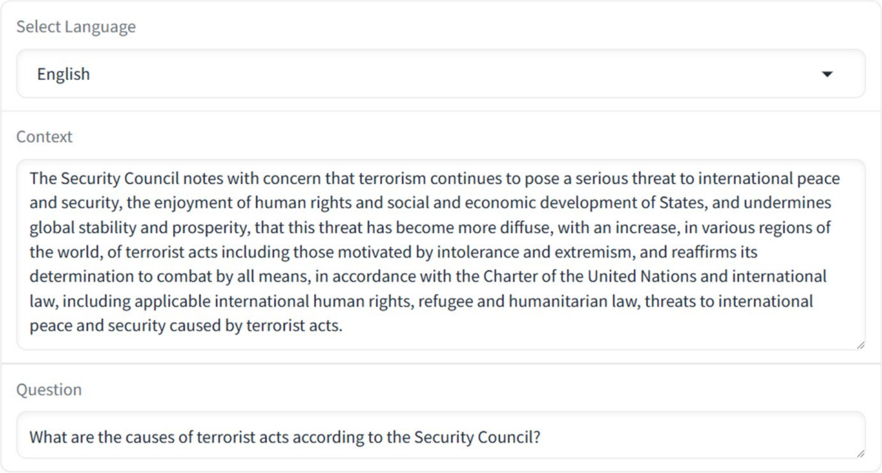
This section presents an illustrative multilingual application of extractive QA using ConfliBERT. Rather than a formal benchmark, the goal is to demonstrate model behavior across real-world, human-translated political texts. Because all UNPC documents are translated by professional UN linguists for formal publication, this corpus provides a rare source of culturally moderated, high-fidelity multilingual text, mitigating many issues associated with automated or informal translation.

To this end, a random sample of United Nations Security Council (UNSC) resolutions in English, Spanish, and Arabic was selected. The context texts in Spanish and Arabic were sourced directly from the UNPC's human-translated resolutions, ensuring high-quality alignment with the English originals. The questions were independently authored in English, Spanish, and Arabic by native-speaking team members with domain expertise. This approach ensured appropriate linguistic, political, and cultural framing in each language, minimizing cross-lingual bias and aligning the evaluation with how political conflict is commonly expressed in each language.

The evaluation focused on three critical topics: human rights, protection of civilians, and terrorism, reflecting common themes in political conflict and violence. The best-performing ConfliBERT models were compared against their respective baseline BERT models.

6.1 English case study

An example was selected from the UNPC corpus, and the following question was posed based on its content: "What are the causes of terrorist acts according to the Security Council?" The English context and corresponding question are shown in Fig. 14. The ConfliBERT-English model identified the correct answer, "motivated by



Select Language

English

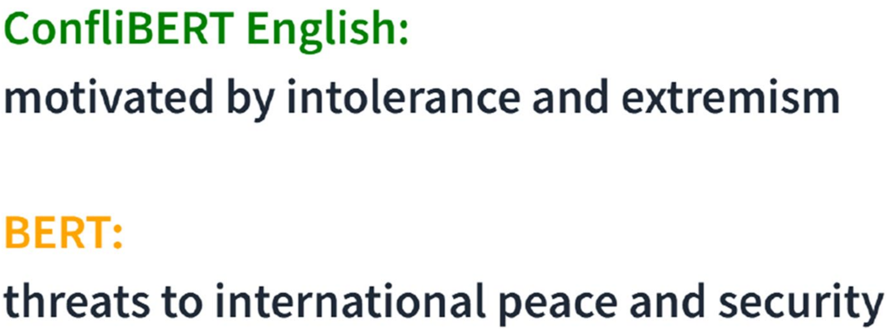
Context

The Security Council notes with concern that terrorism continues to pose a serious threat to international peace and security, the enjoyment of human rights and social and economic development of States, and undermines global stability and prosperity, that this threat has become more diffuse, with an increase, in various regions of the world, of terrorist acts including those motivated by intolerance and extremism, and reaffirms its determination to combat by all means, in accordance with the Charter of the United Nations and international law, including applicable international human rights, refugee and humanitarian law, threats to international peace and security caused by terrorist acts.

Question

What are the causes of terrorist acts according to the Security Council?

Fig. 14 UNPC English Context and Question



ConfliBERT English:

motivated by intolerance and extremism

BERT:

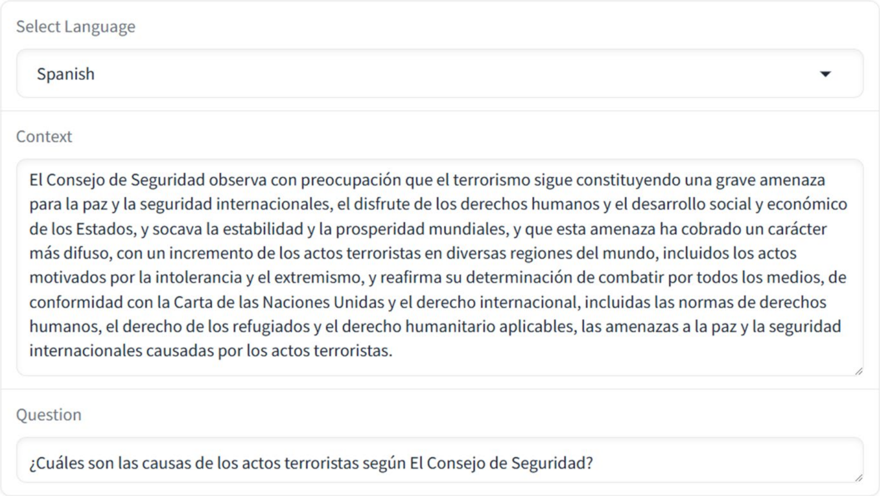
threats to international peace and security

Fig. 15 Predicted Answers from ConfliBERT-English and Baseline BERT

intolerance and extremism” while the baseline BERT model returned the incorrect phrase, “threats to international peace and security” as shown in Fig. 15.

6.2 Spanish case study

The Spanish context text was sourced from the human-translated UNPC corpus. The corresponding question was independently authored in Spanish by a native-speaking team member with domain knowledge, rather than directly translated. The corresponding context and question are shown in Fig. 16. The ConfliBERT-Spanish model returned the correct answer in Spanish, corresponding to the output provided in English. In contrast, the baseline Spanish BERT model returned the same incorrect phrase as in the English example, but rendered in Spanish (Fig. 17).



Select Language

Spanish

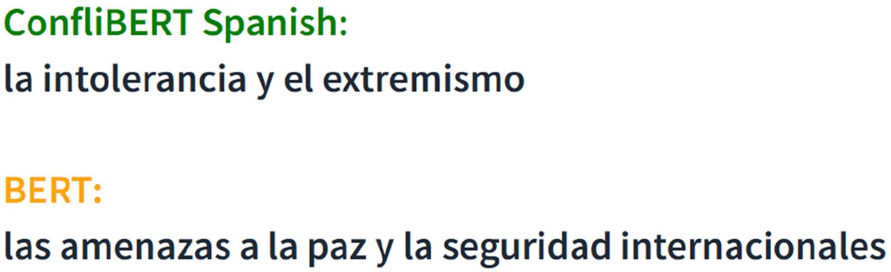
Context

El Consejo de Seguridad observa con preocupación que el terrorismo sigue constituyendo una grave amenaza para la paz y la seguridad internacionales, el disfrute de los derechos humanos y el desarrollo social y económico de los Estados, y socava la estabilidad y la prosperidad mundiales, y que esta amenaza ha cobrado un carácter más difuso, con un incremento de los actos terroristas en diversas regiones del mundo, incluidos los actos motivados por la intolerancia y el extremismo, y reafirma su determinación de combatir por todos los medios, de conformidad con la Carta de las Naciones Unidas y el derecho internacional, incluidas las normas de derechos humanos, el derecho de los refugiados y el derecho humanitario aplicables, las amenazas a la paz y la seguridad internacionales causadas por los actos terroristas.

Question

¿Cuáles son las causas de los actos terroristas según El Consejo de Seguridad?

Fig. 16 UNPC Spanish Context and Question



ConfliBERT Spanish:

la intolerancia y el extremismo

BERT:

las amenazas a la paz y la seguridad internacionales

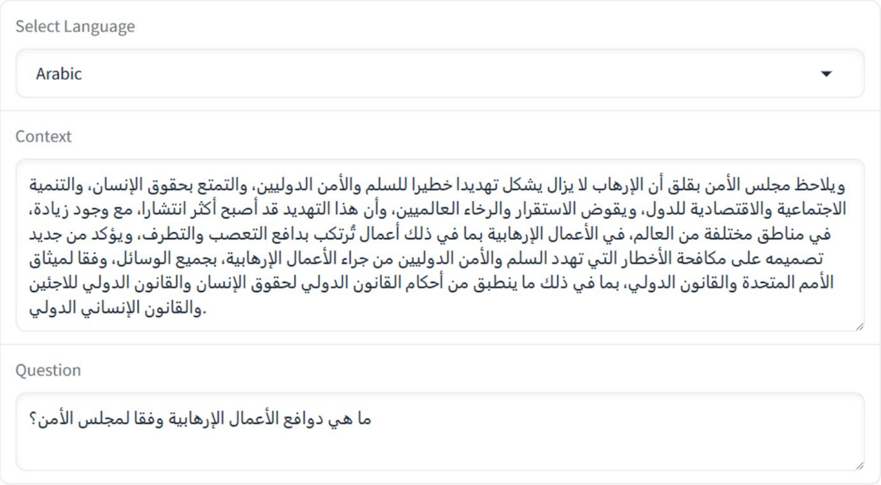
Fig. 17 Predicted Answers from ConfliBERT-Spanish and Baseline BERT

6.3 Arabic case study

The Arabic context text was also sourced from the human-translated UNPC corpus. The question was authored in Arabic by a native-speaking team member familiar with the domain, ensuring contextual alignment. The Arabic context and question are shown in Fig. 18. The ConfliBERT-Arabic model accurately answered the question, with its response corresponding to those provided in English and Spanish. In contrast, the baseline Arabic BERT model produced an incorrect response that was semantically similar to its English and Spanish outputs, but included additional content that was not relevant to the question (Fig. 19).

6.4 Discussion

This illustrative evaluation highlights the advantages of multilingual domain-specific extractive QA models like ConfliBERT. The use of human-translated context texts from the UNPC corpus ensured high-quality input across English, Spanish, and Ara-



Select Language

Arabic

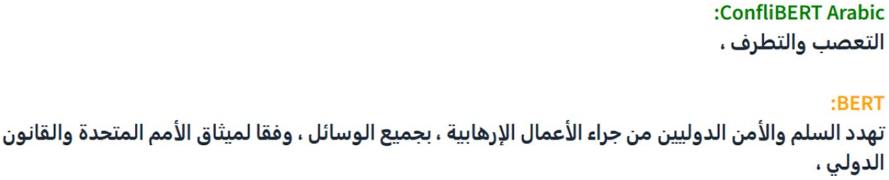
Context

ويلاحظ مجلس الأمن بقلق أن الإرهاب لا يزال يشكل تهديدا خطيرا للسلم والأمن الدوليين، والتمتع بحقوق الإنسان، والتنمية
الاجتماعية والاقتصادية للدول، ويقوض الاستقرار والرخاء العالميين، وأن هذا التهديد قد أصبح أكثر انتشارا، مع وجود زيادة،
في مناطق مختلفة من العالم، في الأعمال الإرهابية بما في ذلك أعمال تُرتكب بدافع التعصب والتطرف، ويؤكد من جديد
تصميمه على مكافحة الأخطار التي تهدد السلم والأمن الدوليين من جراء الأعمال الإرهابية، بجميع الوسائل، وفقا لميثاق
الأمم المتحدة والقانون الدولي، بما في ذلك ما ينطبق من أحكام القانون الدولي لحقوق الإنسان والقانون الدولي للاجئين
والقانون الإنساني الدولي.

Question

ما هي دوافع الأعمال الإرهابية وفقا لمجلس الأمن؟

Fig. 18 UNPC Arabic Context and Question



:ConfliBERT Arabic
التعصب والتطرف ،

:BERT
تهدد السلم والأمن الدوليين من جراء الأعمال الإرهابية ، بجميع الوسائل ، وفقا لميثاق الأمم المتحدة والقانون
الدولي ،

Fig. 19 Predicted Answers from ConfliBERT-Arabic and Baseline BERT

bic, enabling fair comparisons. The manual authoring of questions by native-speaking team members added another layer of precision, ensuring alignment with each language's context and framing.

ConfliBERT consistently provided accurate and aligned answers across languages, demonstrating its ability to process multilingual political texts more effectively than general-purpose baselines. By contrast, the baseline BERT models struggled across all three languages, often repeating similar incorrect answers regardless of language. While this evaluation was based on a small number of representative examples, the results highlight the weaknesses of general-purpose models in specialized domains and underscore the importance of fine-tuned, domain-specific multilingual QA systems for research in political conflict and violence.

7 Conclusion and future work

This study introduced extractive question answering capabilities for ConfliBERT models in English, Spanish, and Arabic, demonstrating effective multilingual adaptation to political and conflict-related domains. An extractive QA methodology was developed for all three languages, with datasets curated to address resource gaps, particularly for Spanish. ConfliBERT models consistently outperformed baseline BERT models across languages, emphasizing the benefits of domain-specific fine-tuning for multilingual QA.

Future work will focus on expanding QA datasets in political and conflict domains, especially for underrepresented languages with limited high-quality resources. Efforts will also include refining dataset translation techniques and exploring closed-book and generative QA systems to extend ConfliBERT's capabilities. Additionally, future directions include extending ConfliBERT to complementary tasks such as summarization, further broadening its applications for multilingual political text analysis.

By enabling high-quality QA over political texts in multiple languages, this work contributes to the broader goal of improving access to conflict-related information for researchers, practitioners, and policy analysts working in diverse global contexts.

8 Limitations

While this study demonstrates promising multilingual QA results in the political domain, several limitations remain. First, while the work focuses on political conflict and violence, the models and evaluations are limited to datasets within this domain, including curated QA resources and UN Security Council texts. Generalization to other domains or genres has not been evaluated, particularly in contexts such as social media or user-generated content. Second, while the models were evaluated using both quantitative and qualitative methods, the qualitative analysis (Section 6) was based on a small number of illustrative examples and does not constitute a comprehensive benchmark.

Third, although questions were authored or reviewed by native speakers with domain expertise, cultural framing and linguistic nuance may still influence alignment between questions and translated context passages. This remains an ongoing challenge in multilingual QA. Finally, while the fine-tuned ConfliBERT models consistently outperformed general-purpose BERT baselines in this domain, their practical utility has not yet been evaluated in downstream applications such as policy analysis, summarization, or multilingual information extraction.

Future work will address these limitations by expanding the evaluation scope, conducting statistical comparisons, incorporating more diverse domains and languages, and testing across applied NLP tasks.

9 Ethical considerations

This research does not involve human participants or the collection of primary data. All analyses are based on publicly available, secondary sources, including curated QA datasets and official documents from the United Nations Parallel Corpus. Where applicable, data usage adheres to licensing and copyright restrictions; as a result, some raw data cannot be publicly released.

To mitigate the risk of bias in multilingual NLP applications, model training and evaluation were conducted on carefully selected domain-relevant corpora, in line with established recommendations for ethical NLP research (Barberá et al. 2021). By developing tools for underrepresented languages in the context of political conflict, this study contributes to broader efforts to expand linguistic equity and data accessibility in political science and computational social science (Magueresse et al. 2020).

This study contributes to advancing more accurate computerized tools to assist scholars and practitioners interested in analyzing violent conflict around the world. As this research highlights, it is crucial to advance computerized tools that minimize error as much as possible, since slight errors can be highly consequential for defense and human security purposes.

Acknowledgements The authors would like to extend their appreciation to the Deanship of Scientific Research at King Saud University for funding this work through Waed Program (W25-99). The authors thank Parker Whitehead for his contributions to the initial codebase used in the early stages of this research. This research was supported in part by NSF award 2311142. This work used Delta at the National Center for Supercomputing Applications (NCSA) / University of Illinois through allocation CIS220162 from the ACCESS program.

Author contributions S.A. developed the scripts and contributed significantly to the manuscript writing and revisions. J.O. contributed to the manuscript writing and conducted analysis. M.A, V.D., L.K., and P.B. provided critical editing, review, and domain-specific expertise, guiding the study's alignment with the target research domain.

Data availability This study used seven datasets. Six are publicly available: SQuAD v1.1, SQAC, XQuAD, MLQA, ARCD, and NewsQA (accessible through reconstruction scripts released by the original authors at https://github.com/Maluuba/newsqa). The Spanish-translated version of NewsQA, created for this study, cannot be shared due to copyright restrictions associated with the original CNN articles.

Declarations

Competing interests The authors declare no competing interests.

References

Alsarra S, Abdeljaber L, Yang W, Zawad N, Khan L, Brandt P, Osorio J, D'Orazio V (2023) Conflibert-arabic: A pre-trained arabic language model for politics, conflicts and violence. In: Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, pp 98–108

Antoun W, Baly F, Hajj H (2020) AraBERT: Transformer-based model for Arabic language understanding. In: Al-Khalifa H, Magdy W, Darwish K, Elsayed T, Mubarak H (eds) Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, pp 9–15. European Language Resource Association, Marseille, France. https://aclanthology.org/2020.osact-1.2/

Artetxe M, Ruder S, Yogatama D (2019) On the cross-lingual transferability of monolingual representations. arXiv:1910.11856

Barberá P, Boydstun AE, Linn S, McMahon R, Nagler J (2021) Automated text classification of news articles: A practical guide. Political Anal 29(1):19–42

Bird S, Loper E (2004) NLTK: The natural language toolkit. In: Proceedings of the ACL Interactive Poster and Demonstration Sessions, pp 214–217. Association for Computational Linguistics, Barcelona, Spain . https://aclanthology.org/P04-3031/

Cañete J, Chaperon G, Fuentes R, Ho J-H, Kang H, Pérez J (2023) Spanish Pre-trained BERT Model and Evaluation Data. https://arxiv.org/abs/2308.02976

Carrino CP, Costa-jussà MR, Fonollosa JAR (2019) Automatic Spanish Translation of the SQuAD Dataset for Multilingual Question Answering

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

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

Gutiérrez-Fandiño A, Armengol-Estapé J, Pàmies M, Llop-Palao J, Silveira-Ocampo J, Carrino CP, Gonzalez-Agirre A, Armentano-Oller C, Rodriguez-Penagos C, Villegas M (2021) Maria: Spanish language models. arXiv:2107.07253

Häffner S, Hofer M, Nagl M, Walterskirchen J (2023) Introducing an interpretable deep learning approach to domain-specific dictionary creation: A use case for conflict prediction. Political Anal 31(4):481–499

Hu Y, Hosseini M, Parolin ES, Osorio J, Khan L, Brandt P, D'Orazio V (2022) ConfliBERT: A pre-trained language model for political conflict and violence. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 5469–5482

Hu J, Ruder S, Siddhant A, Neubig G, Firat O, Johnson M (2020) Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In: International Conference on Machine Learning, pp 4411–4421. PMLR

Jalili Sabet M, Dufter P, Yvon F, Schütze H (2020) SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In: Cohn T, He Y, Liu Y (eds) Findings of the Association for Computational Linguistics: EMNLP 2020, pp 1627–1643. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.findings-emnlp.147

Lee K, Salant S, Kwiatkowski T, Parikh A, Das D, Berant J (2016) Learning recurrent span representations for extractive question answering. arXiv:1611.01436

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

Magueresse A, Carles V, Heetderks E (2020) Low-resource languages: A review of past work and future challenges. CoRR arXiv:2006.07264

Mozannar H, Maamary E, El Hajal K, Hajj H (2019) Neural Arabic question answering. In: Proceedings of the Fourth Arabic Natural Language Processing Workshop, pp 108–118. Association for Computational Linguistics, Florence, Italy. www.aclweb.org/anthology/W19-4612

Osorio J, Reyes A (2017) Supervised Event Coding From Text Written in Spanish: Introducing Eventus ID. Social Sci Comput Rev 35(3):406–416. https://doi.org/10.1177/0894439315625475

Osorio J, Alshammari A, Alatrush N, Heintze D, Converse A, Alsarra S, Khan L, Brandt PT, D'Orazio V (2025) The devil is in the details: Assessing the effects of machine-translation on llm performance in domain-specific texts. Proceed Mach Trans Summit XX 1:315–332

Osorio J, Alsarra S, Converse A, Alshammari A, Heintze D, Khan L, Alatrush N, Brandt PT, D'Orazio V, Zawad N, Billah M (2024) Keep it local: Comparing domain-specific llms in native and machine translated text using parallel corpora on political conflict. In: 2024 2nd International Conference on Foundation and Large Language Models (FLLM), pp 542–552. https://doi.org/10.1109/FLLM63129.2024.10852489

Pires T, Schlinger E, Garrette D (2019) How multilingual is multilingual BERT? In: Korhonen A, Traum D, Màrquez L (eds) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp 4996–5001. Association for Computational Linguistics, Florence, Italy. https://doi.org/10.18653/v1/P19-1493

Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) SQuAD: 100,000+ questions for machine comprehension of text. In: Su J, Duh K, Carreras X (eds) Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp 2383–2392. Association for Computational Linguistics, Austin, Texas. https://doi.org/10.18653/v1/D16-1264

Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D (2021) Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digital Med 4(1):86

Ray PP (2023) Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things Cyber-Phys Syst 3:121–154

Schrodt PA (2009) TABARI. Textual Analysis by Augmented Replacement Instructions. University of Kansas, Lawrence, Kansas. http://eventdata.parusanalytics.com/software.dir/tabari.html

Schrodt PA (2006) Twenty Years of the Kansas Event Data System Project. The Political Methodol 14(1):2–6

Tiedemann J, Aulamo M, Bakshandaeva D, Boggia M, Grönroos S-A, Nieminen T, Raganato A, Scherrer Y, Vazquez R, Virpioja S (2023) Democratizing neural machine translation with OPUS-MT. Language Resource Eval 58:713–755. https://doi.org/10.1007/s10579-023-09704-w

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

Wu B, Cheung AK, Xing J (2021) Learning chinese political formulaic phraseology from a self-built bilingual united nations security council corpus: a pilot study. Babel 67(4):500–521

Yang W, Alsarra S, Abdeljaber L, Zawad N, Delaram Z, Osorio J, Khan L, Brandt P, D'Orazio V (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"

Ziemski M, Junczys-Dowmunt M, Pouliquen B (2016) The united nations parallel corpus v1. 0. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp 3530–3534

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Sultan Alsarra is an Assistant Professor in the Department of Software Engineering at King Saud University. His expertise spans natural language processing, software engineering, and domain-specific large language models, with applications to political conflict, multilingual question answering, healthcare, and translation. He received his Ph.D. in Software Engineering from the University of Texas at Dallas, where his research focused on the design and adaptation of large language models for specialized domains. He has collaborated on multiple projects supported by the U.S. National Science Foundation and the Deanship of Scientific Research at King Saud University, advancing research in multilingual QA systems and large language analysis. Dr. Alsarra's work has led to the development of the ConfliBERT family of models, designed to analyze conflict-related texts across English, Arabic, and Spanish. His research has been published in international venues such as RANLP and ICHI, and he has also contributed to translation-based methods for adapting scarce resources into low-resource languages. He has also been actively involved in collaborative research with teams at the University of Arizona and UT Dallas, focusing on event data, applied AI, and domain-specific NLP systems. He is the recipient of the Best Paper Award at the 2024 SBP-BRiMS conference for his contributions to multilingual QA in political texts. In addition to research, he teaches courses in agile software engineering, ethics in computing, and advanced NLP applications, mentoring students in both applied projects and research-driven initiatives.

Mubarak Alrashoud received the Ph.D. degree in computer science from the Department of Computer Science, Ryerson University, Canada, in 2015. He is currently an Associate Professor and Head of the Department of Software Engineering, College of Computer and Information Sciences, King Saud University. He served for more than a decade in the Royal Saudi Air Defense Forces (RSADF), where he worked in multiple roles, including Ada and C++ programmer, system analyst, software test engineer, and IT management. His research interests include software engineering decision support, optimizing the software release planning process, fuzzy decision making, and optimization in Big Data environments.

Javier Osorio is an Associate Professor in the School of Government and Public Policy at the University of Arizona. His expertise spans political and criminal violence, natural language processing, big data analytics, and quantitative methods in the social sciences. He received his Ph.D. in Political Science from the University of Notre Dame, where his research focused on the dynamics of violence and political conflict in Latin America. Dr. Osorio has led collaborative projects supported by the U.S. National Science Foundation, the Department of Defense–Minerva Initiative, and USAID. He is also the founder of the Academy for Security Analysis, a program dedicated to training practitioners and implementing randomized controlled trials on security in Central America. His work includes advancing supervised event coding methods, cross-linguistic NLP systems, and the integration of machine learning into political science research. His research has been widely published in leading journals such as the American Journal of Political Science, Journal of Peace Research, Journal of Conflict Resolution, and Social Science Computer Review. He has received recognition for his contributions, including the UNODC Best Dissertation Award (2015) and the MPSA Emerging Scholar Best Paper Award (2017). In addition to research, he serves on editorial boards and actively mentors students in computational political science, bridging methodological innovation with substantive research on violence and governance.

Vito D'Orazio is an Associate Professor of Political Science and Data Sciences at West Virginia University. His expertise spans political methodology, conflict forecasting, predictive modeling, and the application of machine learning and natural language processing to political and international relations research. He received his Ph.D. in Political Science from Pennsylvania State University, with concentrations in international relations, methodology, and information science. Prior to joining WVU in 2022, he was on the faculty at the University of Texas at Dallas and served as a postdoctoral researcher in data science at Harvard University's Institute for Quantitative Social Science. Dr. D'Orazio has been a Co-Principal Investigator on the Militarized Interstate Dispute project and the UTD Event Data project, developing systems and models for analyzing large-scale conflict data. His work has been supported by the National Science Foundation and DARPA, and he has published across both political science and data science venues. In addition to research, he teaches courses on political methodology and computational approaches to conflict studies, mentoring students who work at the intersection of social science and artificial intelligence.

Latifur Khan (Fellow, IEEE) is a Professor of Computer Science at the University of Texas at Dallas, where he has been teaching and conducting research since 2000. He received his M.S. and Ph.D. degrees in Computer Science from the University of Southern California in 1996 and 2000. Dr. Khan is an internationally recognized leader in big data analytics, data mining, and large-scale data management, with over 170 publications in journals, conferences, and books. His work has been supported by millions of dollars in federal funding, including multiple National Science Foundation awards, often in collaboration with colleagues in computer science, political science, and public policy. He has served as Program Chair for major international conferences such as IEEE Big Data 2019 and PAKDD 2016, and has delivered tutorials at leading venues including ACM SIGKDD, IJCAI, AAAI, and SDM. Dr. Khan has also served as an associate editor for journals such as ACM Transactions on Internet Technology and IEEE Transactions on Knowledge and Data Engineering. He is an ACM Distinguished Scientist and a Senior Member of IEEE. His research and teaching have established him as a central figure in data science at UT Dallas, where his big data analytics courses attract large enrollments and prepare students for cutting-edge research and industry roles. He is the recipient of numerous honors, including the IEEE Technical Achievement Award for Intelligence and Security Informatics and the Chancellor Award from the President of Bangladesh.

Patrick T. Brandt is a Professor of Political Science at the University of Texas at Dallas. His expertise spans time series analysis, Bayesian statistics, and machine learning methods applied to international relations, political economy, conflict, terrorism, and public opinion. He received his Ph.D. in Political Science from Indiana University, an M.S. in Mathematical Methods in the Social Sciences from Northwestern University, and an A.B. in Government from the College of William and Mary. Dr. Brandt's research develops new statistical models to study change in political and economic events over time, including vector autoregession and count time series approaches. His work has been supported by the National Science Foundation and the Center for Risk and Economic Analysis of Terrorist Events (CREATE). He has published in leading journals of political science and methodology and serves on editorial boards, including the Journal of Conflict Resolution and International Interactions. He is the recipient of multiple recognitions, including the Robert H. Durr Award from the Midwest Political Science Association and the 2024 UT Dallas Recognition of Outstanding Achievements in Research (ROAR) Award. In addition to research, he teaches and mentors students in advanced political methodology, and has led training workshops in Bayesian time series and forecasting models for the social sciences.

Authors and Affiliations

Sultan Alsarra[1] · Mubarak Alrashoud[1] · Javier Osorio[2] · Vito D′Orazio[3] · Latifur Khan[4] · Patrick T. Brandt[5]

✉ Sultan Alsarra
salsarra@ksu.edu.sa

Mubarak Alrashoud
malrashoud@ksu.edu.sa

Javier Osorio
josorio1@arizona.edu

Vito D'Orazio
vito.dorazio@mail.wvu.edu

Latifur Khan
lkhan@utdallas.edu

Patrick T. Brandt
pbrandt@utdallas.edu

[1] Department of Software Engineering, College of Computer and Information Sciences, King Saud University, P.O Box 22452, 11495 Riyadh, Saudi Arabia

[2] School of Government and Public Policy, University of Arizona, 85719 Tucson, United States

[3] Department of Political Science, West Virginia University, 26501 Morgantown, United States

[4] Department of Computer Science, Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, 75080 Richardson, United States

[5] School of Economic, Political and Policy Sciences, University of Texas at Dallas, 75080 Richardson, United States