Extractive Question Answering for Spanish and Arabic Political Text

Sultan Alsarra[1](✉), Parker Whitehead[2], Naif Alatrush[2], Luay Abdeljaber[2], Latifur Khan[2], Javier Osorio[3], Patrick T. Brandt[2], and Vito D'Orazio[4](✉)

[1] King Saud University, Riyadh, Saudi Arabia
salsarra@ksu.edu.sa

[2] The University of Texas at Dallas, Richardson, TX 75080, USA
{parker.whitehead,naif.alatrash,luay.abdeljaber,lkhan, pbrandt}@utdallas.edu

[3] The University of Arizona, Tucson, AZ 85721, USA
josorio1@arizona.edu

[4] West Virginia University, Morgantown, WV 26506, USA
vito.dorazio@mail.wvu.edu

Abstract. This study advances the integration of domain-specific large language models (LLMs) for low-resource languages with applications for question-answering (QA). Leveraging on recent LLMs trained to extract events of political violence and conflict, we introduce ConfliBERT-Arabic and ConfliBERT-Spanish, fine-tuned for extractive QA. Contributions include tailored QA fine-tuning techniques for Arabic and Spanish, curation of five datasets, and a comprehensive performance analysis. These new models provide language and domain-specific enhancements over extant models trained on general corpora. Substantively, these tools allow implementation of high-quality QA about conflict and violence in multiple world regions in their native languages.

Keywords: Large language models · Natural language processing · Question answering · Arabic · Spanish

1 Introduction

The ConfliBERT family of domain-specific large language models (LLMs) are trained on texts about political violence and armed conflict in English [10], Arabic [1], and Spanish [18]. The Spanish and Arabic models have focused on binary classification, named entity recognition (NER), and multi-label classification tasks. This study presents an application about question-answering (QA), and contributes to their growth for low-resource languages.

Researchers [11] have characterized three main types of QA: extractive, open-generative, and closed-generative. Closed-generative and open-generative QA both use the model to generate an answer. In contrast, extractive QA identifies the answer to a question in a given context, without generating new text. For instance, when using a news article about conflict in the Middle East, it can

R. Thomson et al. (Eds.): SBP-BRiMS 2024, LNCS 14972, pp. 144–153, 2024.
https://doi.org/10.1007/978-3-031-72241-7_14

answer questions like which countries are involved or the number of casualties. The output consists of the character indexes in the text that mark the beginning and end of the answer. This is the kind of QA most relevant to the information extraction tasks used in conflict research.

We introduce the ConfliBERT Arabic and Spanish LLMs, fine-tuned for extractive QA for the political violence and conflict domain. Our goal is to elevate their performance in this domain and foster research in QA for low-resource languages. Our work delivers several contributions. First, we implemented fine-tuning techniques tailored to the unique characteristics of Arabic and Spanish for extractive QA. Second, we curated and prepared five datasets, three in Arabic and two in Spanish, providing support for low-resource language processing and expanding NLP accessibility in Spanish by translating the NewsQA dataset from English. Finally, we conducted a comprehensive performance analysis on extractive QA benchmarks, showcasing the enhancements our models offer when compared to base models trained on general corpora.

2 Background and Challenges

BERT [6] opened a new era of NLP capabilities advancing specialized applications including domain-specific tools and improving performance on complex tasks, such as QA. However, domain-specific and QA advances have mostly evolved in parallel without generating integrated tools for domain-specific QA.

2.1 Domain-Specific Developments

Despite its contributions to NLP development, BERT is trained on generic corpora that provide limited leverage for processing text in specialized fields that use technical or specific language. Recognizing this limitation, researchers developed domain-specific adaptations of BERT based on corpora relevant to particular fields. In general, these domain-specific models show higher performance in specialized tasks than the generic BERT [5,15].

In line with domain-specific advances, our research team developed ConfliBERT [10], a BERT-like model specialized on political violence and armed conflict. Results show that ConfliBERT produces much better results than alternative models [8,10], thus reinforcing the trend of enhanced performance by domain-specific models. Based on the initial ConfliBERT architecture in English, this research team advanced multi-lingual extensions of ConfliBERT for Arabic [1] and Spanish [18]. Despite the domain-specific and multi-lingual advances of the ConfliBERT family of models, these tools have not yet been applied to QA.

2.2 Extractive QA Challenges

Extractive QA is widely acknowledged as a complex downstream task for BERT. In this task, the model uses text related to a specific topic as the context, along with a corresponding question. The task is to extract pertinent information from the text to provide an accurate response to the question.

To fine-tune extractive QA, BERT takes the question and context as input, separated by the [SEP] token. The model's output, to the right of the [SEP] token, contains the tokens identified as the 'answer.' Through multiple iterations and training epochs, BERT learns which tokens are relevant for QA.

A key challenge in extractive QA is the limited availability of question-answering datasets, particularly for low-resource languages like Arabic and Spanish. We needed to search for suitable QA datasets, and then modify and process them to fine-tune our models. Discused in detail below, our efforts resulted in the creation of three datasets for Arabic and two for Spanish. These datasets contain the QA pairs used to fine-tuning and test our domain-specific models. We applied extractive QA to both Arabic and Spanish ConfliBERT, comparing the results with their respective base models.

3 Dataset Preparation

For the extractive QA task, we developed scripts tailored to low-resource languages and constructed datasets for ConfliBERT in Arabic and Spanish. The evaluation aimed to measure the effectiveness of our models in real-world applications within the fields of political science and conflict, so it was important to use datasets in this domain.

3.1 Script Development

We developed all scripts from scratch using the HuggingFace library. Our scripts facilitate the QA fine-tuning and evaluation for any language model using Transformers. They have been optimized to handle Arabic and Spanish characteristics, including lower-casing text, removing punctuation, articles, and extra white space. In Arabic, we also removed diacritics like tashkeel and longation, which are phonetic guides unnecessary for the QA task.

To boost the fine-tuning process, we use parallel processing across multiple GPUs to batch multiple fine-tuning jobs by reading JSON arguments. We rely on the National Center for Supercomputing Applications at the University of Illinois with access to 64 CPUs, 248GB of RAM, and 4 A100 GPUs.

All datasets were split into training–used for fine-tuning–and validation–used to assess the model's performance. Both the training and validation sets had five columns: **ID** a unique string for each question-answer pair; **Title** categorizes the context; **Question** the question being asked; **Context** the context in text form; and **Answers** comprising two keys "Answer_Start" (a list of starting indexes for the answers within the context) and "Text" (the raw answer text).

3.2 Spanish Datasets

We use two domain-specific QA datasets for Spanish: NewsQA and SQAC. NewsQA [17] is an extractive QA dataset containing more than 100,000 QA pairs crowd-sourced from CNN articles, with a significant focus on political conflict and violence. The dataset is only available in English, so we translated to Spanish using the Translate Align Retrieve (TAR) [4] method, a proven method for translating extractive QA datasets such as SQUAD [14].

Prepossessing and Cleaning . We accessed the NewsQA data using code made available on Github by Microsoft (github.com/Maluuba/newsqa). Then, we eliminated redundant rows and poor-quality QA pairs, and structured the data into five essential components for our QA tasks: ID, title, context, question, and answers. We stored the formatted data locally in the 'datasetdict' format, which aligns with the extractive QA standards, ensuring consistency and compatibility with other datasets such as SQuAD.

Translation. TAR first uses machine translation to render the question, answers, and context in the target language (Spanish). The original TAR method translates entire contexts as a whole. However, the NewsQA CNN articles are significantly larger than the contexts in SQuAD. Attempting to translate or align them at their full size could introduce inaccuracies and noise. So, we used the NLTK library to split the CNN articles into separate sentences. Then, we employed `opus-mt-en-es`, an English-to-Spanish neural machine translation model for parallel translations. This resulted in a list of translated sentences for each article. The questions were translated directly, without sentence tokenization.

The second TAR step is word alignment. We used `SimAlign`, a novel BERT multilingual approach that aligns source words to their corresponding target words in a sentence. We used the `ArgMax` matching method, as it demonstrated superior performance. By translating the CNN text sentence by sentence, we ensured an equal number of English and Spanish sentences. This allowed us to align only the sentence containing the answer, leading to improved runtime and more efficient memory use. In cases where the answer spanned across multiple sentences, we concatenated and aligned those sentences.

The final TAR step involves retrieving the text. After completing the alignment process, we replaced all English answer text with their corresponding translations in the Spanish text. For answers with empty spaces or gaps, we filled them by highlighting the word with the lowest index next to the word with the highest index, effectively creating our translated answers. Next, we determined the starting index of the answer relative to the translated context, which comprised the concatenation of all sentences from the CNN story. Finally, we converted the dataset back into the datasetdict format, ready for use.

The Spanish Question Answering Corpus (SQAC) [7] is an extractive QA dataset in Spanish. It comprises 6,247 contexts and 18,817 questions, each with 1 to 5 corresponding answers, all sourced from texts originally written in Spanish. These texts were drawn from encyclopedic articles in Spanish Wikipedia, news articles from Wikinews and Newswire, and literary content from AnCora.

To create this dataset we adapted SQuAD v1.1 extractive QA with the help of Native Spanish speakers. Given the dataset's heavy news sources, it contains a substantial number of political QA pairs, providing a valuable resource for evaluating our models in the political domain.

3.3 Arabic Evaluation Datasets

The XQUAD dataset (Cross-Lingual Question Answering Dataset) [2] was developed by Google Deepmind to assess cross-lingual extractive QA performance. It includes 240 paragraphs and 1,190 QA pairs originally sourced from the SquAD v1.1 dataset [14] and translated into multiple languages, including Arabic. For our evaluation, we used XQUAD Arabic, which was prepared for extractive QA tasks by the XTREME benchmark [9]. The Arabic XQUAD dataset contains numerous questions on political topics and subjects.

The MLQA (MultiLingual Question Answering) dataset [12] was developed by Facebook Research to assess cross-lingual QA performance. It includes over 5,000 QA pairs multiple languages, including Arabic. The data are in SQaAD format, have been machine-translated from Wikipedia paragraphs, and have a substantial number of political topics and questions.

The Arabic Reading Comprehension Dataset (ARCD) [13] comprises 1,395 crowd-sourced QA pairs using Arabic Wikipedia articles. The dataset categorizes the answers into numerical answers, such as dates, and non-numerical answers, such as verbs, nouns, or adjective phrases. In cases of noun phrases, they checked for named entities and conducted manual verification. Additionally, the authors manually labeled the dataset for synonyms, world knowledge, syntactic variation, multiple-sentence reasoning, and ambiguity to facilitate question reasoning. We split the dataset into train/test sets, with 702 questions drawn from 78 articles in the test set. Many of these articles focus on political subjects and topics.

4 Experimental Setup

We fine-tuned eight models in Spanish and five in Arabic. For Spanish, the ConfliBERT models were base-multilingual-cased, base-multilingual-uncased, BETO-cased, and BETO-uncased. Each of these has a corresponding BERT model. For Arabic, the fine-tuned ConfliBERT models were Arabic-v2-araBERT, Arabic-v2-multilingual-uncased, and Arabic-v2-Scratch. The comparable base models were BERT-base-araBERT and BERT-base-multilingual-uncased.

We assessed our model's performance using two standard metrics: Exact Match and F1 Score. For fine-tuning, we maintained consistency by using the same hyperparameters across all models. These hyperparameters follow best practices for fine-tuning BERT-based models, as outlined in the original BERT paper [6]. We ran experiments over 5 epochs with 5 different seeds, using a batch size of 8 and a learning rate of 5e-5, a maximum answer length of 100 and a maximum context length of 384.

5 Results and Analysis

The domain-specific ConfliBERT models consistently outperform more generic BERT models for these extractive QA tasks. The QA datasets had questions about political violence, wars, elections, protests, etc., and thus the ConfliBERT models were able to make use of their domain-focused training.

5.1 ConfliBERT-Spanish QA

The results of the Spanish models fine-tuned for extractive QA are shown in Table 1, with Table 1a, showing the average performance across the two datasets. The ConfliBERT-Spanish model outperformed the comparable base BERT model, and often by a substantial margin. The F1 Score for ConfliBERT BETO-Uncased is nearly 7 points higher than its BERT counterpart. Overall, the best model was the ConfliBERT BETO-Cased.

Table 1. Results for Spanish

Model Name		(a) Extractive AQ		(b) News QA		(c) SQAC	
		F1 Score	Exact Match	F1 Score	Exact Match	F1 Score	Exact Match
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

5.2 ConfliBERT-Arabic QA

Table 2 shows the results for the Arabic models, with Table 2a showing the average across datasets. Once again, the ConfliBERT models performed better than their BERT counterparts in every case. For each of the three datasets and for each metric, the best performing model was ConfliBERT AraBERT.

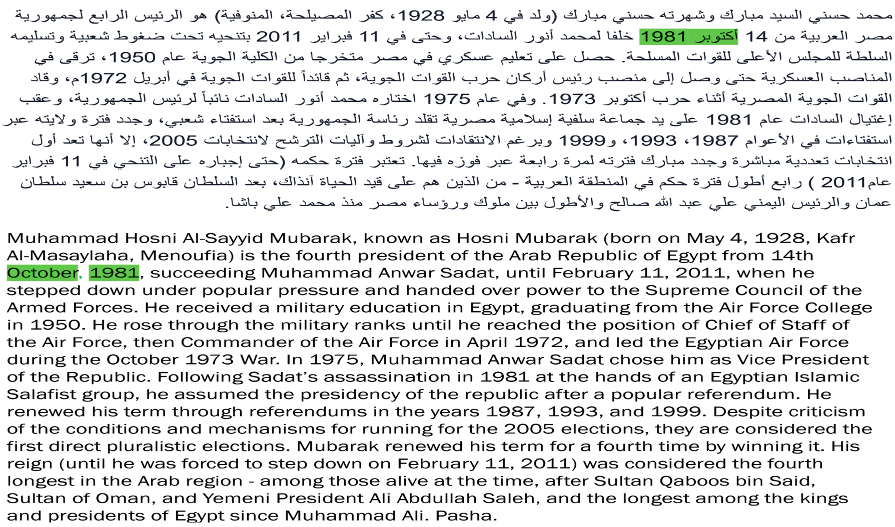
Table 2. Results for Arabic.

Model Name		(a) Extractive QA		(b) MLQA		(c) XQUAD		(d) ARCD	
		F1 Score	Exact Match	F1 Score	Exact Match	F1 Score	Exact Match	F1 Score	Exact Match
ConfliBERT Arabic-v2	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

5.3 Evaluation of Answers

Here, we present examples of QA pairs used to assess our models. We conducted a comparative analysis, comparing the results of the best-performing ConfliBERT-Arabic models against the best-performing base BERT models. We also provide a brief comparison with the responses generated by ChatGPT [16].

The initial set of questions focused on the context related to the former president of Egypt, Hosni Mubarak. The first question posed in Arabic was Q1: “When did Hosni Mubarak take over the reins of power in Egypt?” We supplied the models with the context for Q1, and they generated responses based on this information. The context and answer for Q1 in Arabic, along with its equivalent translation in English, are provided in Fig. 1. The correct answer to Q1 is October 1981, with its position highlighted in the context.



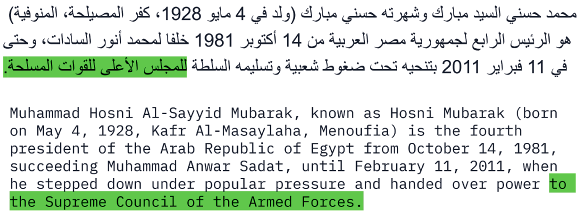
محمد حسني السيد مبارك وشهرته حسني مبارك (ولد في 4 مايو 1928، كفر المصيلحة، المنوفية) هو الرئيس الرابع لجمهورية
مصر العربية من 14 أكتوبر 1981 خلفا لمحمد أنور السادات، وحتى في 11 فبراير 2011 بتنحيه تحت ضغوط شعبية وتسليمه
السلطة للمجلس الأعلى للقوات المسلحة. حصل على تعليم عسكري في مصر متخرجا من الكلية الجوية 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 Sadat chose him 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. 1. Q1 Context and Highlighted Answer

The ConfliBERT-Arabic model predicted the answer as “1981,” correctly identifying the year but omitting the month. In contrast, base BERT incorrectly predicted “Year 1950” as the answer, corresponding to the year when Hosni Mubarak graduated from the Air Force College, as indicated in the context.

After introducing additional context, we prompted the models with a more intricate question concerning Hosni Mubarak: “To whom did Hosni Mubarak hand power after the 2011 protests?” The correct answer for Q2 is “to the Supreme Council of the Armed Forces.” Fig. 2 provides the context, along with the highlighted answer for Q2. The ConfliBERT-Arabic models accurately predicted the answers, providing the exact response. Conversely, the base BERT model generated an incorrect answer, offering the date when Mubarak handed over power as “11 2022 February.”



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. 2. Q2 Context and Highlighted Answer.

Overall, ConfliBERT-Arabic extractive QA models consistently outperformed base BERT, providing accurate responses to most of the questions. Notably, base BERT faced challenges, particularly with political questions. Also, our ConfliBERT-Arabic demonstrated proficient handling and accurate representation of date formats and numbers in answers. In contrast, base BERT struggled, presenting dates and numbers in incorrect orders and formats.

We also conducted a comparative analysis with ChatGPT. For instance, we prompted ChatGPT with "Answer questions based on the following text:" and then provided the same Q1 context in Arabic as shown in Fig. 1. We then asked ChatGPT the same Q1 question, "When did Hosni Mubarak take over the reins of power in Egypt?" It provided the answer in Arabic, and we then requested it to be translated into English. ChatGPT provided a correct answer (October 14, 1981) but introduced inaccuracies by stating that Hosni Mubarak assumed power after the *resignation* of President Mohamed Anwar Sadat. In reality, Sadat did not resign, he was assassinated.

In our experiments with ChatGPT, particularly in the analysis of political text, we observed a tendency to infer extra details not present in the text, potentially leading to incorrect answers and biased responses based on question wording. These issues highlight the importance of caution when relying on models like ChatGPT for detailed and accurate analysis of political texts. In contrast, our domain-specific ConfliBERT models provide answers directly from the provided text and context without introducing extraneous details or generating nonsensical responses. This is an extremely important attribute for researchers who want to use these models for information extraction tasks.

6 Conclusion and Future Work

We introduced extractive QA for ConfliBERT-Arabic and ConfliBERT-Spanish models. This involved crafting an extractive QA methodology from the ground up for the unique aspects of Arabic and Spanish. We curated extractive QA datasets and undertook the translation of an English dataset to Spanish, contributing to the needs of low-resource languages in NLP. Our evaluation compared model performance against the base BERT models showing how our Arabic and Spanish models excelled especially in the domains of politics and violence.

Our future research develops more QA datasets for the political and conflict domains. We also plan to enhance our dataset translation methodologies and introduce closed-generative QA. A significant challenge in extractive QA for low-resource languages is the dearth of non-machine-translated datasets. Crafting specialized datasets for Arabic and Spanish will prove instrumental to advance the use of low-resource languages in NLP. Furthermore, refining translation techniques will extend linguistic resources for low-resource languages by enabling the translation of existing QA datasets. Our future ConfliBERT tasks aspire to create closed generative QA and expand into areas like summarization.

7 Ethical Considerations

The tools generated here provide NLP resources tailored for languages with low resources to reduce bias in academia and the policy sector. This study relies exclusively on secondary sources of information such as news reports and does not engage with human subjects to gather information from primary sources. To address concerns of biased ML inputs, it complies to select corpora and training data [3]. Due to copyright protecting the original sources, we cannot share the raw data.

Acknowledgements. This research was supported in part by NSF awards 1737978, 2039542, 1828467, 1931541, 2311142, and 1906630; ONR awards 17-1-2995 and 20-1-2738; and Army Research Office Contract No. W911NF2110032. This work used Delta at NCSA/University of Illinois through allocation CIS220162 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by NSF grants 2138259, 2138286, 2138307, 2137603, and 2138296. This material is based upon High Performance Computing (HPC) resources supported by the University of Arizona TRIF, UITS, and Research, Innovation, and Impact (RII).

References

1. Alsarra, S., et al.: 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 (2023)
2. Artetxe, M., Ruder, S., Yogatama, D.: On the cross-lingual transferability of monolingual representations. arXiv preprint arXiv:1910.11856 (2019)
3. Barberá, P., Boydstun, A.E., Linn, S., McMahon, R., Nagler, J.: Automated text classification of news articles: a practical guide. Polit. Anal. **29**(1), 19–42 (2021)
4. Carrino, C.P., Costa-juss, M.R., Fonollosa, J.A.R.: Automatic Spanish translation of the squad dataset for multilingual question answering (2019)
5. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I.: Legal-bert: the muppets straight out of law school. arXiv preprint arXiv:2010.02559 (2020)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

7. Gutiérrez-Fandiño, A., et al.: Maria: Spanish language models. arXiv preprint arXiv:2107.07253 (2021)
8. Häffner, S., Hofer, M., Nagl, M., Walterskirchen, J.: Introducing an interpretable deep learning approach to domain-specific dictionary creation: a use case for conflict prediction. Polit. Anal. **31**(4), 481–499 (2023)
9. Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., Johnson, M.: Xtreme: a massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In: International Conference on Machine Learning, pp. 4411–4421. PMLR (2020)
10. Hu, Y., et al.: 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 (2022)
11. Lee, K., Salant, S., Kwiatkowski, T., Parikh, A., Das, D., Berant, J.: Learning recurrent span representations for extractive question answering. arXiv preprint arXiv:1611.01436 (2016)
12. Lewis, P., Oğuz, B., Rinott, R., Riedel, S., Schwenk, H.: MLQA: evaluating cross-lingual extractive question answering. arXiv preprint arXiv:1910.07475 (2019)
13. Mozannar, H., Maamary, E., El Hajal, K., Hajj, H.: Neural Arabic question answering. In: Proceedings of the Fourth Arabic Natural Language Processing Workshop, pp. 108–118. Association for Computational Linguistics, Florence (2019). www.aclweb.org/anthology/W19-4612
14. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. arXiv e-prints arXiv:1606.05250 (2016)
15. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D.: Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Dig. Med. **4**(1), 86 (2021)
16. Ray, P.P.: Chatgpt: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber-Phys. Syst. **3**, 121–154 (2023)
17. Trischler, A., et al.: Newsqa: a machine comprehension dataset. In: Proceedings of the 2nd Workshop on Representation Learning for NLP, pp. 191–200 (2017)
18. Yang, W., et al.: 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" (2023)