The 2nd International Conference on Foundation and Large Language Models (FLLM2024)

2024 2nd International Conference on Foundation and Large Language Models (FLLM) | 979-8-3503-5479-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/FLLM63129.2024.10852489

Keep it Local: Comparing Domain-Specific LLMs in Native and Machine Translated Text using Parallel Corpora on Political Conflict

Javier Osorio
School of Government and Public Policy
University of Arizona
Tucson, United States
josorio1@arizona.edu

Sultan Alsarra
Department of Software Engineering
King Saud University
Riyadh, Saudi Arabia
salsarra@ksu.edu.sa

Amber Converse
Department of Linguistics
University of Arizona
Tucson, United States
aconverse@arizona.edu

Afraa Alshammari
Department of Computer Science
University of Texas - Dallas
Dallas, United States
afraa.alshammari@utdallas.edu

Dagmar Heintze
School of Economic, Political and Policy Sciences
University of Texas - Dallas
Dallas, United States
dagmar.heintze@utdallas.edu

Latifur Khan
Department of Computer Science
University of Texas - Dallas
Dallas, United States
lkhan@utdallas.edu

Naif Alatrush
Department of Computer Science
University of Texas - Dallas
Dallas, United States
naif.alatrash@utdallas.edu

Patrick T. Brandt
School of Economic, Political and Policy Sciences
University of Texas - Dallas
Dallas, United States
pbrandt@utdallas.edu

Vito D'Orazio
Department of Political Science
West Virginia University)
Morgantown, United States
vito.dorazio@mail.wvu.edu

Niamat Zawad
Department of Computer Science
University of Texas - Dallas
Dallas, United States
niamat.zawad@utdallas.edu

Mahrusa Billah
Department of Computer Science
University of Texas - Dallas
Dallas, United States
mahrusa.billah@utdallas.edu

***Abstract*—The dynamics of political conflict and cooperation require powerful computerized tools capable of effectively tracking security threats and cooperation around the world. This study compares the performance of domain-specific Large Language Models (LLMs) against generically-trained LLMs in binary and multi-class classification using native text in English, Spanish, and Arabic, and their corresponding machine translations. This endeavor yields four key contributions. 1) We present and make available a novel database of annotations using a multi-lingual parallel corpus from the United Nations. 2) Using various metrics, we assess the quality of different machine translation tools. 3) Our results indicate that the ConfliBERT family of LLMs, a set of domain-specific models tailored for political conflict, outperform generically-trained LLMs in English, Spanish, and Arabic in both binary and multi-class tasks. 4) We also disentangle the heterogeneous effects of machine translation on LLM performance in different languages. Overall, results reveal the comparative advantage of native-language domain-specific LLMs specialized on political conflict to understand the dynamics of violence and cooperation worldwide using native text. Our multi-lingual ConfliBERT LLMs provide critical cyber-infrastructure enabling scholars and government agencies use their local languages and information to foster safer, more stable political environments.**

***Index Terms*—Multilingual LLMs, machine translation, political conflict, United Nations.**

I. Introduction

Contemporary political conflict and cooperation are characterized by rapidly changing dynamics beyond the traditional military and diplomatic interactions of nation states. Emerging from local insurgencies, terrorists, criminal organizations, ethnic conflict, human and drug trafficking, social unrest, and piracy, among others, a great variety of conflict incidents and cooperation opportunities involve non-state armed actors that need to be analyzed. Tracking, understanding, and mitigating these complex conflict and cooperation processes requires

This research was supported in part by NSF awards DMS-1737978, DGE-2039542, OAC-1828467, OAC-2311142, and DGE-1906630, ONR awards N00014-17-1-2995 and N00014-20-1-2738, 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). The PIs also thank the amazing annotators based in the US, Mexico, and Lebanon who contributed to this study.

979-8-3503-5479-9/24/$31.00 ©2024 IEEE

leveraging powerful computerized approaches capable to identify such incidents in an effective and timely manner.

To address these challenges, multidisciplinary endeavors combining computer scientists and political scientists have advanced computerized applications to study conflict. Early efforts such as ICEWS [1] relied on rule-based coders such as TABARI [2] and PETRARCH2 [3] to code conflict events using large dictionaries applying the CAMEO [4] ontology. However, the rapidity changing conflict dynamics demanded considerable costs to update these dictionaries, quickly relegating them to obsolescence. Moreover, the rigidity of rule-based coders often lacked the capacity to analyze complex unstructured text describing political conflict processes.

Recent Large Language Models (LLMs) developments such as BERT [5] shattered the limitations preventing rule-based coders from effectively processing unstructured text and promised broad possibilities to study conflict. Despite their ground-breaking contributions, generically pre-trained LLMs struggle when processing domain-specific tasks related to political conflict [6]. Most conflict coding efforts exclusively rely on English-language and just a few use machine translation [7], thus overlooking valuable information in foreign languages and introducing considerable coverage bias [8].

This study makes four significant contributions to advance conflict research using computerized tools. First, the paper presents an annotated database in English, Spanish, and Arabic to comparatively analyze the performance of domain-specific and generically-trained LLMs to study political conflict. The study leverages multi-lingual parallel data from the United Nations (UN) and presents a large set of high-quality annotations relevant to political conflict and cooperation. Secondly, this study explicitly evaluates the use of machine translation in conflict research, gauging the quality of different machine translation tools using various metrics. Thirdly, we compare the performance of the ConfliBERT family models [6], [9], [10], a set of domain-specific LLMs tailored for political conflict, against several generically-trained models in English, Spanish, and Arabic. This evaluation encompasses both binary and multi-class classification of machine-translated data and native texts. Lastly, the study disentangles the effects of machine translation to better understand variations in LLM performance across machine-translated and native texts.

II. Related Work

Political violence and cooperation research is a focal point for scholars and security professionals focused on tracking, analyzing, and predicting social unrest, political violence, and armed conflicts worldwide [1], [11]–[14]. In political science, conflict analysis examines a broad spectrum of interactions of government entities, non-state actors, and civilians. Studying political confrontation and cooperation encompasses a broad range of behaviors such as protests, riots, crackdowns, insurgencies, civil wars, terrorism, human rights abuses, genocides, forced displacements, conventional and unconventional wars, nuclear deterrence, peacekeeping efforts, diplomatic tensions and tensions, international aid, and collaborative initiatives.

Conflict scholars have long been applying computerized methods to study conflict processes around the world. Initial developments in this field used rule-based coders such as TABARI [2] and PETRARCH2 [3] that employed large dictionaries of actors and actions to generate conflict data. Early rule-based coders worked exclusively on English text which prevented them from processing data in foreign languages. This shortcoming motivated later efforts to generate rule-based coders in Spanish [15] and Arabic [16], and rule-translation efforts [17]. The CAMEO ontology [4] became the dominant schema for event coding efforts such as ICEWS [1], the Phoenix Data Set [18], and TERRIER [19]. Recently, the PLOVER [14] political event classification superseded CAMEO with a more succinct set of action categories that facilitate the coding process. A central limitation of rule-based coders consisted on the costs, labor, and time required to update the dictionaries on a regular basis. There were efforts to automatically update coding dictionaries [20]–[22] or translate them into non-English languages [17], yet the rapid changing conflict processes made them perennially outdated.

To address these challenges, researchers developed automated tools, particularly transformer-based pre-trained language models (PLM) [5], [23], [24]. Leveraging self-supervision on vast amounts of unlabeled text, these models reduce the necessity for dictionaries or extensive manual annotation through transfer learning.

A recent contribution is ConfliBERT [6], a domain-specific model specifically designed for classifying political conflict and cooperation in (English) texts that significantly decreases the need for extensive expert human annotation. Its development involves two primary steps: (1) training a BERT-based LLM on a domain-specific corpus focused on political violence, and (2) evaluating the model across various downstream tasks. Later developments extended ConfliBERT's multi-lingual capabilities to Arabic and Spanish languages: ConfliBERT-Arabic [25] and ConfliBERT-Spanish [26]. These models enhance conflict research in their respective languages, reflecting growing interests to expand linguistic resources to directly study conflict in foreign locations using native sources.

Social scientists are increasingly developing, adopting, and adapting computer science and computational linguists tools to study conflict [27]–[31]. Machine learning has been used to facilitate both data generation [7], [32], [33], conflict analysis [34]–[38], and improving conflict prediction abilities about changes in the levels of political violence [39]–[41]. The primary focus across these applications is on texts written in English language. Low-resource languages, such as Arabic and Spanish, are frequently underrepresented and require extensive adjustments to prevent under-performance, which decreases their attractiveness and accuracy for usage [42].

Adding NLP analysis tools from low-resource languages and regions to better understand conflict presents new challenges to state-of-the-art NLP models. For a domain- and language-specific LLM problem like the one considered in this study, most studies exploit machine translation from low resource languages to English to capture the information from low-resource languages and reduce their variability [1], [7], [43]. Liu et al. [44] implement this technique by using off-the-shelf and fine-tuning approaches to translate data from French and Chinese into English, and evaluate multiple encoders. However, this machine-translation approach circumvents rather than addresses the challenge of developing domain and language specific NLP tools to analyze native text.

III. Data

A. The United Nations Parallel Corpus

The empirical foundations of this study rely on the United Nations Parallel Corpus (UNPC) [45], a large collection of official United Nations (UN) documents from 1990–2014. The UNPC contains 86,307 documents professionally and manually translated by the UN Department for General Assembly and Conference Management (DGACM) Translation Services into all six official UN languages (English, Spanish, Arabic, French, Russian, and Chinese). These documents are aligned at the sentence-level and contain a total of 11,365,709 fully aligned sentences. This study uses a random sample of 7,800 sentences from United Nations Security Council (UNSC) resolutions in English (EN), Spanish (ES), and Arabic (AR) related to three key topics (human rights, the protection of civilians, and terrorism). This yields a highly relevant corpus for the domain of political conflict and violence, thus constituting a suitable case for testing the leverage of domain and language-specific ConfliBERT LLMs and a direct comparison to generically-trained LLMs in a single language with a machine translation step. Our corpus is generally comprised of relatively short sentences with an average length 27 words in English, 31 words in Spanish, and 24 words in Arabic.

B. Annotation Procedure

To prepare the 7,800 UNPC sentences for analysis, annotations were made by 12 human coders with domain expertise in political science and international relations, and bi-lingual skills in either English-Spanish or English-Arabic. The annotators were given randomly sampled sentences to classifying according to 1) their relevance or non-relevance, and 2) the QuadClass categories of Verbal/Material-Conflict/Cooperation in Table I using Label Studio [46], an annotation interface capable of processing text in multiple languages.[1]

TABLE I
QuadClass Categories

	Cooperation	Conflict
Verbal	Agree Consult Support Concede	Demand Disapprove Reject Threaten
Material	Cooperate Aid Retreat Investigate	Protest Crime Sanction Mobilize Coerce Assault

The annotation procedure consisted of seven steps. 1) All coders underwent a rigorous training process to gain familiarity with the codebook. 2) Each week, coding teams received a random sample of the corpus consisting of about 300 sentences aligned across languages. 3) A pair or triplet of human coders conducted a first round of blind annotations for each sentence, assigning a QuadClass category to the relevant sentences and classifying it as relevant or non-relevant if no QuadClass category was identifiable. 4) Coders conducted a non-blind revision round on each sentence. This allowed coders to compare their decisions to those of other coders. 5) The annotations for which there is 100% agreement between coders are considered as Gold Standard Records (GSR). 6) Coders conducted a third non-blind revision round focusing on the remaining sentences with less than 100% agreement. Having multiple coders looking at the same sentence multiple times contributed to improving their inter-coder reliability. 7) For sentences with less than 100% agreement, a coder was assigned to make the final decision to assign the best classification as GSR. Sentences where a final decision was not made were excluded from the final dataset for downstream tasks, as were complex sentences that received more than one QuadClass category label.

C. Binary and Multi-Class Annotations

To illustrate the substantive content of each QuadClass category, each quadrant of Table I includes a set of action types that correspond to the PLOVER ontology [14]. The annotations are a multi-class task indicating whether a sentence represents an instance of Verbal Conflict, Verbal Cooperation, Material Conflict, or Material Cooperation.[2] The manual annotation had high inter-coder reliability (average 92.0% agreement).

Coders annotated a total of 11,493 sentences. Figure 1 presents the distribution of the binary classification where coders identified 52.4% of the data as not relevant, and the other 47.6% as relevant sentences. Figure 2 shows the distribution of the multi-class QuadClass annotations. Of all sentences used for training in the multi-class classification task, coders identified 53.2% as not relevant, 13.7% are Material Conflict, 13.2% as Material Cooperation, 8.3% as Verbal Conflict, and 11.6% as Verbal Cooperation[3]. Of the relevant sentences identified in the binary task, human coders classified 29.4% of the data as Material Conflict, 28.2% correspond to Material Cooperation, 17.7% as Verbal Conflict, and 24.8% correspond to Verbal Cooperation. To avoid the problems of unbalanced data, all the experiments conducted in this study use balanced databases capped at the least common denominator across categories in each classification tasks.

Table II illustrates the annotators' sentences for binary and multi-class classifications. The first binary example shows the

[1] The researchers thank Label Studio for granting access to their app.

[2] Since barely 3% of the sentences had more than one label, we treat this as a multi-class classification task rather than multi-label classification.

[3] The discrepancy in not relevant percentages between Figures 1 and 2 originates from multi-label sentences. We excluded sentences with multiple labels from the multi-class classification task. Therefore, the proportion of not relevant sentences is slightly higher in Figure 2

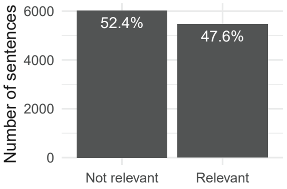

Fig. 1. Binary Annotations

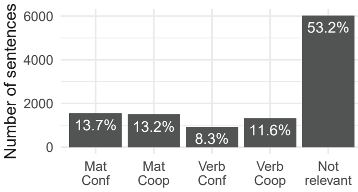

Fig. 2. Multi-Class Annotations

relevance classification and the remaining provide examples for each of the QuadClass categories.

TABLE II
ANNOTATION EXAMPLES

Binary	Relevant	Suicide bombings have also become a trend. También se han hecho frecuentes los atentados suicidas. كما أن التفجيرات الانتحارية أصبحت أمرا مألوفا.
	Not Relevant	I ran towards my cow and untied it. Fui corriendo a desatar a mi vaca. فأسرعت إلى بقرتي وفككت قيدها.

Multi-Class	Material Conflict	Killing of wounded enemy servicemen Asesinatos de combatientes enemigos heridos قتل الجرحى من جنود العدو
	Material Cooperation	Establishment of the integrated command centre Establecimiento del centro de mando integrado ألف - إنشاء مركز القيادة المتكاملة
	Verbal Conflict	Colonel Reis and Mr. Malik had a heated verbal exchange. El Coronel Reis y el Sr. Malik tuvieron una discusión acalorada. ودار جدال شفوي محتدم بين العقيد ريس والسيد مالك.
	Verbal Cooperation	Countering terrorist narratives and violent extremism Lucha contra la retórica terrorista y el extremismo violento مناهضة الخطاب الإرهابي والتطرف المقترن بالعنف

IV. Machine Translation

A. Machine-Translation Tools

Rather than using native language text to study conflict in foreign locations, conflict scholars often rely on machine-translated text suitable for English-based NLP tools [7]. This seems to be a cost-effective strategy as it takes advantage of a rich set of NLP tools available in English without the burden of using or developing NLP tools in foreign languages. However, this approach seldom rests on a systematic assessment of the machine translation quality and often neglects errors derived from distorted translation. To address this limitation, this section evaluates the quality of the output of different machine translation tools. First, we use various tools to translate Spanish (ES) and Arabic (AR) native UNPC sentences into English language (EN). Second, we assess the quality of machine translation using a variety of metrics. Third, we evaluate the performance of different English-language models on two downstream tasks using the English machine-translated text.

The UNPC data constitutes an invaluable resource pairing professional manually translated sentences across languages for a single corpora, allowing one to compare the leverage of different common machine translation tools. However, it is plausible to expect variations in the capacity of different machine translation tools to accurately preserve contextual meaning and accuracy in the translation output.

To evaluate the quality of the machine translation, we translated the entire sample of Arabic and Spanish sentences into English. We use four machine translation tools: 1) Google Translate [47], which is commonly used for machine translations in academic research [7], 2) DeepL API [48], a high-quality neural machine translation service recognized for its superior performance, providing accurate and contextually attentive translations, 3) Deep Learning translator, a Python library package that abstracts the complexities of API usage and introduces a simple interface for translation services [49], and 4) OPUS [50], a Hugging Face library providing state-of-the-art NLP PLM. For the latter, we use the Helsinki-NLP/opus-mt-ar-en and Helsinki-NLP/opus-mt-es-en models since they were specifically trained to translate between Arabic and English and Spanish to English.

All four translation tools differ in their functionality, which may affect the translation output. For Google Translate, we relied on Google's subscription-based Cloud Translation service. Google Translate relies on a transformer-based neural network system and an RNN decoder. In addition to changing the system architecture, Google Translate relies on new training data from embedding-based model web crawls, using a data miner that prioritizes precision over recall [51]. For DeepL, we used the advanced subscription-based translation GUI. DeepL relies on artificial neural networks based partially on the transformers architecture. However, the network topology differs from other commonly used transformers-based tools, which improves DeepL's performance. The model is trained on specifically collected training data and relies on web crawlers that detect online translations and conduct quality assessments [52]. The Deep Learning translator package is a Python package that uses multiple translators, such as Google Translate, Mymemory Translator, DeeplTranslator, QcriTranslator, Linguee Translator, PONS Translator, Yandex Translator, Microsoft Translator, ChatGpt Translator, Papago Translator, Libre Translator, TencentTranslator, and BaiduTranslator. The Python tool includes multiple translators and supports a variety of source text formats. It relies on an API server, thereby facilitating fast and larger batch-size translations [49]. For the analysis in this paper, we relied on google translate within the Deep Learning Python package. In contrast to the subscription-based Google Cloud translator, the free Google translator provided in the package can be considered to be less reliable, and subject to throttling and breaking [53], [54]. Our final machine translation tool, OPUS, equally relies on a transformer-based neural machine translation architecture. The model is trained on freely available parallel corpora that were collected for the OPUS bitext repository [55]. While all translation tools use different approaches to conducting the machine translations, they all rely on a transformers-based architecture and different training data. Deep Learning adds additional complexity by combining different translation tools in one Python package. Consequently, we expect the machine translation output to differ. To better understand the differences in machine-translation outputs, we conduct additional quality assessments.

B. Machine-Translation Quality Metrics

To assess the translation quality, we use four different metrics to compare the machine translated English text (from the original ES and AR sentences) to the native UNPC English ground truth. The first metric is BLEU (Bilingual Evaluation Understudy) [56], which calculates the precision of n-grams (sequences of n words) in the machine-generated text compared to the ground truth. As the most rigid metric, BLEU does not assess the contextual correctness of translations but rather evaluates whether each word from the source text was correctly translated into a corresponding target text word. The second metric is SacreBLEU [57], a variation of BLEU that addresses tokenization and normalization matters to ensure the evaluation is comparable and consistent across various systems. This allows SacreBLEU for some translation flexibility [58]. The third metric is METEOR [59], which uses explicit ordering to create a word alignment between the translated text and the ground truth, and calculates the similarity scores for them. Finally, BERT-score is a metric that employs a BERT model [5] to calculate the similarity between the machine-translated text and the ground truth based on high-level semantic features. BERT-score provides the most flexible translation metric, permitting a contextually correct translation and the use of synonyms without restricting the assessment to the correctness of word-by-word translation. The scores from these metrics range from 0 to 1, with a higher score representing a higher level of similarity of the translated text to the ground truth [60].

Table III presents the translation quality assessment of the machine translation from Spanish and Arabic into English. The metrics listed in the Table are ordered from the strictest (BLEU) to the most flexible scale (BERTscore). DeepL achieves the best results across all tools for both languages with a top BERTscore of 0.9668 for the Spanish to English translation and 0.9638 for the Arabic to English text. The second-best performance is OPUS, with a performance that significantly challenges DeepL in all four quality metrics. These findings demonstrate the general role of deep learning in the performance of text translations. Results from Table III also reveal that Google Translate yields the lowest quality across metrics in both languages. This questions the validity of the approach used in other studies relying on Google [7]. Despite variations of quality scores across different machine-translation tools, these metrics show that there is not much substantial variation between translation tools. This is an important factor, specially considering the monetary costs of using DeepL and Deep Learning, *vis-à-vis* the free use of Google Translate and OPUS.

TABLE III
MACHINE TRANSLATION QUALITY

Lang	Metric	Google	DeepL	Deep Learning	OPUS
ES-EN	BLEU	0.4071	0.4467	0.4147	0.4071
	SacreBLEU	0.4611	0.4990	0.4707	0.4611
	METEOR	0.6907	0.7164	0.6965	0.6907
	BERTScore	0.9611	**0.9668**	0.9639	0.9611
AR-EN	BLEU	0.3747	0.4327	0.3792	0.3747
	SacreBLEU	0.4271	0.4859	0.4349	0.4271
	METEOR	0.6739	0.7125	0.6765	0.6739
	BERTScore	0.9553	**0.9639**	0.9571	0.9553

Bold font indicates top results.

TABLE IV
DOMAIN-SPECIFIC AND GENERIC MODELS USING MACHINE TRANSLATED TEXT INTO ENGLISH

Model	ES to EN		AR to EN	
	Binary	MCC	Binary	MCC
ConfliBERT-Cont-Case	0.9213	**0.6305**	0.9165	0.6644
ConfliBERT-Cont-Unc	0.9200	0.6266	0.9140	0.6637
ConfliBERT-Scr-Case	0.9240	0.6239	0.9153	0.6638
ConfliBERT-Scr-Unc	**0.9256****	0.6282	**0.9176*****	**0.6682**
mBERT-Case-fine	0.9139	0.6007	0.9125	0.6299
mBERT-Unc-fine	0.9142	0.5961	0.8944	0.6335
BERT-Case-fine	0.9202	0.6191	0.9132	0.6588
BERT-Unc-fine	0.9226	0.6277	0.9137	**0.6660**
Electra-disc-fine	0.9205	**0.6301**	0.9133	0.6622
RoBERTa	0.9179	0.6235	0.9089	0.6607

Machine translated text using DeepL. Average F1 reported for binary and average macro F1 for multi-class classification (MCC). Bold font indicates top results. Statistical significance $^{*}p<0.1$, $^{**}p<0.05$, $^{***}p<0.01$.

C. LLM Performance Using Machine-Translated Text

We use DeepL, the top performing translation tool, to compare the operation of different LLMs on machine translated text for both the binary and mutli-class classification. The evaluation uses ConfliBERT-EN [6] and generically trained models including BERT [5], multilingual BERT, Electra [61], and RoBERTa [62]. The assessment considers different models including ConfliBERT English with Continual training using cased text (ConfliBERT-Cont-Case), as well as uncased text (ConfliBERT-Cont-Unc), and their corresponding versions with training from scratch, ConfliBERT-Scr-Case and ConfliBERT-Scr-Unc, respectively. As baseline models, we use base BERT with cased text (BERT-Case-fine) and uncased text (BERT-Unc-fine), multilingual BERT in its cased (mBERT-Case-fine) and uncased versions (mBERT-Unc-fine), as well as Electra-disc-fine and RoBERTa.

Table IV presents the average F-1 score results of these evaluations for 10 seeds and 5 epochs for each model with a 70-15-15 split for training, developing, and testing. For the defined tasks, the ConfliBERT family models generally perform better than generic PLM baselines. For the Spanish to English translations, results for binary classification show that ConfliBERT-Src-Unc has the best performance and is highly statistically significant compared to BERT-Unc-fine, the closest generic model competitor. For multi-class classification

(MCC) task in Spanish, ConfliBERT-Cont-Case has the highest average F-1 score, but it is not statistically different from Electra, the closest performing generic model.

For Arabic to English translation, Table IV shows the classification performance of ConfliBERT-Scr-Unc has the top performance in the binary classification, and is statistically superior to BERT-Unc-fine, its closest generic competitor. Results from the MCC task of QuadClass classification using English text translated from Arabic indicate that ConfliBERT-Cont-Unc performs better than BERT-Unc-fine, its closest competition from the generic family of models, yet the difference is not statistically significant.

Overall, the results indicate the comparative advantage of using ConfliBERT, a domain-specific model about political conflict and cooperation. ConfliBERT EN offers a clearly superior performance over generic models for classifying relevant and not relevant sentences in the binary task using translated text. However, the lack of statistical significance in the multi-class classification task indicates that ConfliBERT's comparative advantage is less clear when identifying instances of material and verbal conflict and cooperation using machine translated text from both Spanish and Arabic into English.

V. Native Language Experiments

This section compares the performance of the language-specific ConfliBERT models and generic PLMs on the binary and multi-class downstream tasks using native language UNPC sentences in English, Spanish, and Arabic. All reported models are evaluated based on 10 seeds with 5 epochs using a 70-15-15 data split for training, developing, and testing.

A. Experimental setup

Starting with ConfliBERT English (EN), Hu et al. [6] offered a considerable improvement on a variety of NLP tasks focused on political violence and conflict when compared to Google's BERT. Based on its outstanding performance, Häffner et al. [63] consider ConfliBERT as the state-of-the-art tool for processing conflict event data. Subsequent developments extended ConfliBERT's multilingual capacity in ConfliBERT Spanish (ES) [9] and ConfliBERT Arabic (AR) [10]. These are pre-trained on a large collection of documents specialized on political conflict and violence in their respective native languages. In this study, we compare the performance of the ConfliBERT models and other generic models in English, Spanish, and Arabic using fine-tuned models for both the binary and multi-class classification tasks on their respective languages using the annotated UNPC sentences.

Based on the UN parallel corpus, we conduct two key comparisons: across languages and within languages. First, we assess the performance of different models across languages using the different versions of ConfliBERT in English, Spanish, and Arabic compared to Google's BERT in its English and multilingual versions [32, cf.]. We compare the performance of ConfliBERT with Continual training from multilingual BERT using cased text (ConfliBERT-Cont-Case) and uncased text (ConfliBERT-Cont-Unc) in English, Spanish, and Arabic languages. In addition, we use ConfliBERT from scratch and uncased versions (ConfliBERT-Scr-Unc). ConfliBERT from scratch is only available in English (ConfliBERT-Scr-Case).

We then compare the performance of domain-specific ConfliBERT models to other powerful models trained on generic text. For the baseline, we fine-tuned Google's multilingual BERT base with cased text (mBERT-Case-fine) and uncased text (mBERT-Unc-fine) in English, Spanish, and Arabic languages. In addition, we use BERT base with cased text (BERT-Case-fine) and uncased text (BERT-Unc-fine), which are only available in English. This experimental set up leverages the multilingual character of the UNPC as ground truth to compare the the performance of ConfliBERT models across languages with the Google BERT models as a baseline.

The second assessment dimension focuses on comparing different models within languages. Here we compare the above mentioned models to other broadly used models developed exclusively in each of their corresponding languages. For English, we include Electra [61], which uses replaced token detection instead of BERT's masking as pre-training method. We also rely on RoBERTa [62], which uses a robustly optimized pre-training method. For Spanish, we use BETO [64], a BERT-based model pre-trained exclusively on Spanish text. The assessment includes BETO's cased (BETO-Case) and uncased version (BETO-Unc). In addition, we rely on the continual version of ConfliBERT based on BETO originally developed by Wang et al. [9] in its cased (ConfliBERT-BETO-Case) and uncased (ConfliBERT-BETO-Unc) form. Finally, for Arabic, we use AraBERT [65], a BERT-like model specifically pre-trained with Arabic text. Since the Arabic language does not have capitalized letters, we only use the uncased form of this model. This set up enables the comparison of domain-specific and generic-purpose models within each language.

B. Binary Classification Results

Table V presents the average F-1 score for the binary classification task derived from running each model with 10 seeds and 5 epochs. The top section in Table V presents the ConfliBERT family models across languages. The middle section reports the performance of Google BERT models. Finally, the bottom section presents the results of other generically pre-trained models for specific languages.

Table V shows that the ConfliBERT family models perform better than or as good as generically trained models for classifying relevant or not relevant sentences in their native languages. For the binary classification task in English, ConfliBERT-Scr-Unc performs as well as BERT-Case-fine. Results for the binary task in Spanish language indicate that ConfliBERT-BETO-Unc also performs as well as BETO-Case, the generically-trained Spanish model with the top performance. Finally, Arabic results show that ConfliBERT-AraBERT yields the top results for binary classification using the original UNPC Arabic corpus. This improvement in performance is statistically significant when compared to AraBERT, the closest generic Arabic model competitor.

TABLE V
BINARY CLASSIFICATION USING DOMAIN-SPECIFIC AND GENERIC MODELS ON NATIVE LANGUAGES

Model	EN	ES	AR
ConfliBERT-Cont-Case	0.9375	0.9139	0.8992
ConfliBERT-Cont-Unc	0.9384	0.9150	0.9068
ConfliBERT-Scr-Case	0.9373		
ConfliBERT-Scr-Unc	**0.9392**		0.8976
ConfliBERT-AraBERT			**0.9075*****
ConfliBERT-BETO-Case		0.9146	
ConfliBERT-BETO-Unc		**0.9166**	
mBERT-Case-fine	0.9319	0.9114	0.8826
mBERT-Unc-fine	0.9319	0.9116	0.8890
BERT-Case-fine	**0.9392**		
BERT-Unc-fine	0.9376		
Electra-dis-fine	0.9340		
RoBERTa-fine	0.9286		
BETO-Case-fine		**0.9173**	
BETO-Unc-fine		0.9139	
AraBERT			0.8970

Average F1 reported. Bold font indicates top results.
Statistical significance * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

TABLE VI
MULTI-CLASS CLASSIFICATION CLASSIFICATION USING DOMAIN-SPECIFIC AND GENERIC MODELS ON NATIVE LANGUAGES

Model	EN	ES	AR
ConfliBERT-Cont-Case	0.6569	0.6296	0.6149
ConfliBERT-Cont-Unc	0.6482	0.6288	**0.6291*****
ConfliBERT-Scr-Case	**0.6612*****		
ConfliBERT-Scr-Unc	0.6556		0.5803
ConfliBERT-AraBERT			0.6275
ConfliBERT-BETO-Case		**0.6409**	
ConfliBERT-BETO-Unc		0.6293	
mBERT-Case-fine	0.6161	0.5959	0.5614
mBERT-Unc-fine	0.6222	0.6064	0.5549
BERT-Case-fine	0.6308		
BERT-Unc-fine	0.6362		
Electra-dis-fine	0.6500		
RoBERTa-fine	0.6511		
BETO-Case-fine		**0.6375**	
BETO-Unc-fine		0.6154	
AraBERT			0.5096

Average macro F1 reported. Bold font indicates top results.
Statistical significance * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

In general, the binary classification results show the advantage of relying on language-specific models specifically designed for analyzing text on the domain of political conflict in their native languages. This is particularly the case for processing native text in Arabic. Another key characteristic of the results derived from using domain-specific native models to process native languages is the high F1 scores reached by the top performing models. The high level of performance of these domain-specific models provide effective computerized assistance to researchers and practitioners in identifying valuable information in the massive collection of UN documents with a high degree of accuracy.

C. Multi-Class Classification Results

Table VI reports the results of using domain-specific and generically-trained models to classify incidents of verbal and material conflict and cooperation in UNPC documents across languages. The performance metric reported is the average macro F1 score, which is calculated as the mean for all four individual F1 scores associated with each QuadClass category.

Table VI confirms the superiority of ConfliBERT models for multi-class classification in their native languages. The QuadClass classification in English indicates that ConfliBERT-Scr-Case is the model that provides has the best performance to identify different QuadClass incidents in the UNPC. The macro F1 boost derived from this model is statistically significant compared to RoBERTa, the generically-trained model with the closest performance. Results for Spanish show that ConfliBERT-BETO-Case has slightly better performance than BETO-Case, the two top performing models for QuadClass classification in Spanish. However, this difference is not statistically sificant. Finally, the multi-class classification task conducted on Arabic text using Arabic-specific models indicates that ConfliBERT-Cont-Unc is the best tool for classifying QuadClass instances. The performance of this domain-specific model is statistically significant when compared to mBERT-Case, the top generically-trained Arabic model.

Another characteristic that stands out in Table VI is the relatively lower macro F1 across models as compared to the higher F1 scores of the binary classification in Table V. It seems that identifying instances of material and verbal conflict and cooperation is substantially more difficult than classifying relevant information. This lower performance may be related to the relatively small number of annotations in each QuadClass category discussed in the III Data section.

VI. DIFFERENTIAL PERFORMANCE BETWEEN MACHINE-TRANSLATED AND NATIVE-LANGUAGE TEXT

The results from the machine translation and experimental sections reveal a counter-intuitive finding. The original expectation motivating this study was that domain-specific native LLMs would perform better when processing native text than English-based LLMs applied to text machine-translated into English. However, at first glance, the results do not seem to support this expectation.

To further evaluate these seemingly puzzling results, Table VII compares the top performing models from the machine translation analysis using DeepL in Table IV and the best native models using native text for both the binary and multi-class tasks derived from Tables V and VI, respectively. The analysis then uses a t-test to calculate the difference between the translated and the native language experiments for each task (binary/multi-class), in each language (ES/AR), and for each type of text (machine-translated/native). The last column in Table VII presents the difference of means for each pair of scores with their corresponding statistical significance. As in the original tables, binary task results are F1 scores and those of the multi-class task represent macro F1 scores.

The differential performance in Table VII generally indicates that analyzing machine translated text using English-based models yields marginally better performance than using

TABLE VII
DIFFERENTIAL PERFORMANCE

Task		Text	Best Model	Score	Diff
Binary	ES	Trans.	ConfliBERT-Scr-Unc	0.9256	0.0090***
		Native	ConfliBERT-BETO-Unc	0.9166	
	AR	Trans.	ConfliBERT-Scr-Unc	0.9176	0.0101***
		Native	ConfliBERT-AraBERT	0.9075	
MCC	ES	Trans.	ConfliBERT-Cont-Case	0.6305	-0.0104***
		Native	ConfliBERT-BETO-Case	0.6409	
	AR	Trans.	ConfliBERT-Scr-Unc	0.6682	0.0391***
		Native	ConfliBERT-Cont-Unc	0.6291	

Results from binary classification represent average F1 scores, while results from multi-class classification (MCC) are average macro F1 scores. Statistical significance * p<0.1, ** p<0.05, *** p<0.01.

native models to process native text. Although the magnitude of the difference between pairs of models is very small, the performance improvement is statistically significant across models. While it appears preferable to use machine-translated text and English-based domain-specific ConfliBERT models over native language text and native models, further analysis reveals important shortcomings from such approach.

Given the grammatical and syntactical differences, distinct ConfliBERT models pre-trained in English, Spanish, or Arabic, may perform differently due to the inherent particularities of their corresponding languages. Linguistically, English relies on more concise and succinct grammatical and syntactical structures than Spanish and Arabic, that possess greater morphological complexity and allow for a more flexible word order [42]. Both Spanish and Arabic are pro-noun drop languages that permit more intricate verb conjugations and syntactical constructs than English. Spanish sentences tend to be longer with a greater number of subordinate clauses. Rhetorically, short sentences are perceived as monotonous or redundant in Spanish [66]. In addition, the diacritics in Spanish (*acentos* and *vergulilla*) and Arabic (*Harakat*) provide a richer alphabet in those languages than in English. *Abjads*, such as Arabic, further commonly include homonyms that can only be distinguished in context [42]. In consequence, Spanish and Arabic thrive with longer, more complex, and fluid grammatical and syntactical arrangements.

In an effort to understand the apparent improvement of machine translated text over the native data, we explore the effects of machine translation. To do so, we first disentangle the difference in the number of words generated by machine translation into English when compared to the number of words in the original native language (ES/AR). We then disentangle the machine translation effects by identifying instances in which the machine-translation tool increased or decreased the word count at the sentence level. Figure 3 uses Locally Weighted Scatterplot Smoothing (LOWESS) regression [67] to visualize trends of word count increase or decrease caused by DeepL machine translation for each language.

Trends in Figure 3 reveal that machine translation from Spanish and Arabic into English induces heterogeneous distortions in the data. DeepL translation both increases and decreases sentence-level word counts that disproportionately affect different languages, which are likely to affect the performance of English-based in different ways. The machine translation tool increases the number of words in some instances, while reducing the number of words in other sentences, but the overall net effect of DeepL translation from Spanish into English results in a more succinct corpus, while the net effect of DeepL translation from Arabic is a more verbose corups.

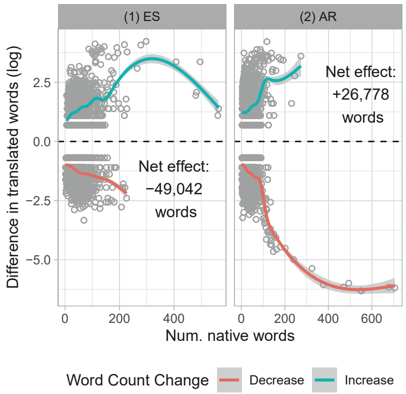

Fig. 3. Translation Effects on Word Count

As the left panel in Figure 3 shows, DeepL has a net word reduction effect on translations from Spanish to English. Although DeepL tends to further elongate a handful of long sentences in Spanish (top left trend in Figure 3), there are more short Spanish sentences that get even shorter as this translation tool turns them into English (lower left trend in Figure 3). In the aggregate, Spanish DeepL translation reduces the word count by -49,042 words, which corresponds to a -13.83% reduction from the total word count in the native Spanish corpus. In contrast, the right panel in Figure 3 indicates that DeepL increases the word count when translating from Arabic into English. Although there are a few long sentences in Arabic that DeepL translates into a more succinct version (lower right trend in Figure 3), most native sentences in Arabic experience a word count increase in their English translation (top right trend in Figure 3). In total, Arabic DeepL translation increases the total number of words by 26,778, which corresponds to a 9.75% increase in the word count from native Arabic text.

The word count increase and decrease effects are consequential for the quality of machine translation into English. Figure 4 reports the different translation quality scores (BLEU, ScareBLEU, METEOR, and BERTScore) for each translated sentence from both Spanish and Arabic. To facilitate the data interpretation, the plots use LOEWSS regression to represent general trends. In general, the plots of the BLEU, ScareBLEU, METEO sentence-level scores indicate that the quality of the translation is poor, while BERTScore, the more forgiving metric, shows a high quality output. Most importantly, Figure 4 show the impact of Spanish word reduction and Arabic word increase on the quality of the English translation. As the Spanish panel shows, providing a more succinct corpus is

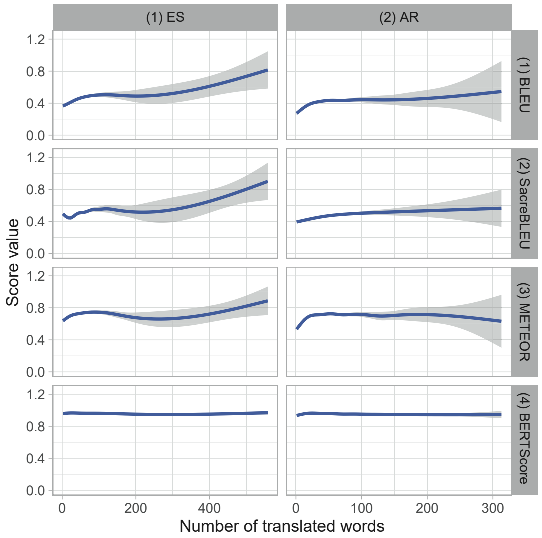

Fig. 4. Quality of Spanish/Arabic to English Translation Scores

rewarded in English as the different quality metrics increase. In contrast, the right panel shows that the quality scores penalize the data augmentation in the English translation from Arabic. According to the LOWESS trends, the quality scores mostly stagnate and the METEOR score even shows a slight decline as the number of translated words increases.

Due to grammatical and syntactical variations, distinct ConfliBERT models developed for Spanish or Arabic may exhibit different performance levels to those of ConfliBERT English given the intrinsic linguistic characteristics of each language. In consequence, theword count reduction induced by DeepL in the Spanish to English translation may artificially improve the quality of the translation and the ConfliBERT EN performance given that English language favors more succinct text. Consequently, using ConfliBERT EN on machine translated text shows better performance than the output of ConfliBERT ES processing native Spanish text. This seems to be the case for the binary classification task in Spanish. In contrast, the DeepL word count increase in the Arabic to English translation provides a more verbose corpus, thus reducing the quality of the translation. Yet, this data increase seems to provide more linguistic elements that artificially improve the ConfliBERT EN classification performance.

VII. Conclusions

This study highlights the comparative advantage of analyzing native-language texts in English, Spanish, and Arabic using domain-specific ConfliBERT LLMs to generate high-quality data on conflict processes to understand political violence and cooperation worldwide. This paper advances the research frontiers in computer science and political science in different ways. The study presents a large collection of high-quality cross-lingual annotations from United Nations data, thus providing a valuable resource to analyze political conflict and cooperation across languages. This aligned database allows scholars to compare the performance of different models using the same informational content in different languages. Future versions of this data will include annotations using the PLOVER [14] ontology, Named Entity Recognition, and Question and Answering.

Evaluating the output of different machine translation tools reveals that scholars should assess in a systematic and transparent way the quality of the machine translation resources they use. Prominent coders such as POLECAT [7] or ICEWS [1] rely on low quality machine translation or are not transparent about their translation tools and output, thus casting doubt about the quality of the data they generate. Using different quality assessment metrics of varying degrees of rigidity, our evaluation indicates that DeepL provides the most accurate translations for both Spanish to English and Arabic to English. However, the detailed analysis of the machine-translated texts reveals heterogeneous word count increase and decrease effects that have consequences for LLM performance. Future works should analyze in a more granular way the distortions caused by machine translation tools and their effects on LLM performance. Also, future work should consider assessing other translation tools such as MS Azure [68] or more recent resources like Claude 3 Opus [69] or ChatGPT [70], which have shown good results in resource-poor languages [71].

The primary contribution of this study is the performance comparison of domain-specific LLMs against generically-trained models using machine-translated data and native texts in English, Spanish, and Arabic. Results show the power of the ConfliBERT family models to generate high-quality data on conflict and cooperation using native-language texts in English, Spanish, and Arabic. This analysis requires significant computational resources and extensive GPUs for fine-tuning the models and conducting the multi-lingual comparisons across models. While this research relied on large computing resources [72], researchers can access localized versions of these resources to advance their own research using the methodology and tools discussed.

By making our annotated databases and our multi-lingual ConfliBERT LLMs publicly available,[4] we contribute to advancing NLP tool for resource-pool languages. Moreover, by providing this public critical cyber-infrastructure, our research tools enable scholars, security practitioners, and government agencies in a large number of English, Spanish, and Arabic speaking countries leverage their local languages and information sources to foster safer, more stable political environments.

VIII. Ethics Statement

This research uses United Nations data as second-hand accounts of political conflict, but does not involve human research subjects. By generating high-quality data on political conflict, this study contributes to understanding the causes and consequences of conflict. By developing native-language NLP tools, this study contributes to enriching low-resource languages, and promotes diversity in STEM.

[4] The annotations and replication materials are available on GitHub at: https://github.com/javierosorio/keep_it_local

References

[1] E. Boschee, J. Lautenschlager, S. O'Brien, S. Shellman, and J. Starz, "ICEWS Weekly Event Data," Harvard Dataverse, 2018.

[2] P. A. Schrodt, "TABARI. Textual Analysis by Augmented Replacement Instructions," Lawrence, Kansas, 2009. [Online]. Available: http://eventdata.parusanalytics.com/software.dir/tabari.html

[3] C. Norris, P. Schrodt, and J. Beieler, "Petrarch2: Another event coding program," *Journal of Open Source Software*, vol. 2, no. 9, p. 133, 2017.

[4] D. J. Gerner, P. A. Schrodt, O. Yilmaz, and R. Abu-Jabr, "Conflict and mediation event observations (cameo): A new event data framework for the analysis of foreign policy interactions," *International Studies Association, New Orleans*, 2002.

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

[6] Y. Hu, M. Hosseini, E. S. Parolin, J. Osorio, L. Khan, P. Brandt, and V. D'Orazio, "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*, 2022, pp. 5469–5482.

[7] A. Halterman, B. E. Bagozzi, A. Beger, P. Schrodt, and G. Scraborough, "Plover and polecat: A new political event ontology and dataset," Apr 2023. [Online]. Available: osf.io/preprints/socarxiv/rm5dw

[8] L. C. Windsor, "Bias in Text Analysis for International Relations Research," *Global Studies Quarterly*, vol. 2, no. 3, p. ksac021, Jul. 2022. [Online]. Available: https://doi.org/10.1093/isagsq/ksac021

[9] W. Yang, S. Alsarra, L. Abdeljaber, N. Zawad, Z. Delaram, J. Osorio, L. Khan, P. T. Brandt, and V. D'Orazio, "Conflibert-spanish: A pre-trained spanish language model for political conflict and violence," in *2023 7th IEEE Congress on Information Science and Technology (CiSt)*, 2023, pp. 287–292.

[10] S. Alsarra, L. Abdeljaber, W. Yang, N. Zawad, L. Khan, P. Brandt, J. Osorio, and V. D'Orazio, "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*, R. Mitkov and G. Angelova, Eds. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria, Sep. 2023, pp. 98–108. [Online]. Available: https://aclanthology.org/2023.ranlp-1.11

[11] J. A. Goldstone, R. H. Bates, D. L. Epstein, T. R. Gurr, M. B. Lustik, M. G. Marshall, J. Ulfelder, and M. Woodward, "A global model for forecasting political instability," *American journal of political science*, vol. 54, no. 1, pp. 190–208, 2010.

[12] E. G. Rød, T. Gåsste, and H. Hegre, "A review and comparison of conflict early warning systems," *International Journal of Forecasting*, vol. 40, no. 1, pp. 96–112, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207023000018

[13] H. Hegre, M. Allansson, M. Basedau, M. Colaresi, M. Croicu, H. Fjelde, F. Hoyles, L. Hultman, S. Högbladh, R. Jansen, N. Mouhleb, S. A. Muhammad, D. Nilsson, H. M. Nygård, G. Olafsdottir, K. Petrova, D. Randahl, E. G. Rød, G. Schneider, N. von Uexkull, and J. Vestby, "ViEWS: A political violence early-warning system," *Journal of Peace Research*, vol. 56, no. 2, pp. 155–174, Mar. 2019, publisher: SAGE Publications Ltd. [Online]. Available: https://doi.org/10.1177/0022343319823860

[14] Open Event Data Alliance, "Political language ontology for verifiable event records," https://github.com/openeventdata/PLOVER, 2018, accessed: 2022-10-01.

[15] J. Osorio and A. Reyes, "Supervised Event Coding From Text Written in Spanish: Introducing Eventus ID," *Social Science Computer Review*, vol. 35, no. 3, pp. 406–416, 2017. [Online]. Available: http://ssc.sagepub.com/content/early/2016/01/07/0894439315625475.abstract

[16] J. Osorio, A. Reyes, A. Beltrán, and A. Ahmadzai, "Supervised event coding from text written in Arabic: Introducing Hadath," in *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*. Marseille, France: European Language Resources Association (ELRA), May 2020, pp. 49–56. [Online]. Available: https://www.aclweb.org/anthology/2020.aespen-1.9

[17] J. Osorio, V. Pavon, S. Salam, J. Holmes, P. T. Brandt, and L. Khan, "Translating CAMEO verbs for automated coding of event data," *International Interactions*, vol. 45, no. 6, pp. 1049–1064, 2019.

[18] S. Salam, P. Brandt, J. Holmes, and L. Khan, "Distributed framework for political event coding in real-time," *Proceedings of the 2018 conference on Electrical Engineering and Computer Science (EECS)*, 2018.

[19] C. Grant, A. Halterman, J. Irvine, Y. Liang, and K. Jabr, "TERRIER," Dec. 2017, publisher: OSF. [Online]. Available: https://osf.io/4m2u7/

[20] M. Solaimani, S. Salam, L. Khan, P. T. Brandt, and V. D'Orazio, "Apart: Automatic political actor recommendation in real-time," in *Social, Cultural, and Behavioral Modeling*, D. Lee, Y.-R. Lin, N. Osgood, and R. Thomson, Eds. Cham: Springer International Publishing, 2017, pp. 342–348.

[21] ——, "Repair: Recommend political actors in real-time from news websites," *Proceedings of the International Conference on Big Data (Big Data)*, pp. 1333–1340, 2017.

[22] B. J. Radford, "Automated dictionary generation for political eventcoding," *Political Science Research and Methods*, vol. 9, no. 1, pp. 157–171, Jan. 2021.

[23] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[24] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[25] S. Alsarra, L. Abdeljaber, W. Yang, N. Zawad, L. Khan, P. Brandt, J. Osorio, and V. D'Orazio, "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*, 2023, pp. 98–108.

[26] W. Yang, S. Alsarra, L. Abdeljaber, N. Zawad, Z. Delaram, J. Osorio, L. Khan, P. T. Brandt, and V. D'Orazio, "ConfliBERT-Spanish: A pre-trained Spanish language model for political conflict and violence," in *2023 7th IEEE Congress on Information Science and Technology (CiSt)*. IEEE, 2023, pp. 287–292.

[27] P. A. Schrodt, "Pattern recognition of international event sequences: A machine learning approach," in *Artificial Intelligence And International Politics*. Routledge, 1991, num Pages: 25.

[28] E. Deutschmann, J. Lorenz, and L. G. Nardin, *Advancing Conflict Research Through Computational Approaches*. Cham: Springer International Publishing, 2020, pp. 1–19. [Online]. Available: https://doi.org/10.1007/978-3-030-29333-8_1

[29] J. Grimmer, M. E. Roberts, and B. M. Stewart, "Machine learning for social science: An agnostic approach," *Annual Review of Political Science*, vol. 24, pp. 395–419, 2021.

[30] ——, *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press, 2022.

[31] L.-E. Cederman and L. Girardin, "Computational approaches to conflict research from modeling and data to computational diplomacy," *Journal of Computational Science*, vol. 72, p. 102112, 07 2023.

[32] A. Halterman, P. A. Schrodt, A. Beger, B. E. Bagozzi, and G. I. Scarborough, "Creating custom event data without dictionaries: A bag-of-tricks," *arXiv preprint arXiv:2304.01331*, 2023.

[33] A. Hürriyetoğlu, H. Tanev, V. Zavarella, J. Piskorski, R. Yeniterzi, D. Yuret, and A. Villavicencio, "Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021): Workshop and Shared Task Report," in *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, 2021, pp. 1–9.

[34] R. A. Blair and N. Sambanis, "Forecasting civil wars: Theory and structure in an age of "big data" and machine learning," *Journal of conflict resolution*, vol. 64, no. 10, pp. 1885–1915, 2020.

[35] T. Anders, "Territorial control in civil wars: Theory and measurement using machine learning," *Journal of Peace Research*, vol. 57, no. 6, pp. 701–714, Nov. 2020. [Online]. Available: https://doi.org/10.1177/0022343320959687

[36] J. Osorio and A. Beltrán, "Enhancing the Detection of Criminal Organizations in Mexico using ML and NLP," *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, 2020, glasgow, Scotland.

[37] J. Osorio, M. Mohamed, V. Pavon, and B.-O. Susan, "Mapping Violent Presence of Armed Actors," *Advances in Cartography in GIScience of the International Cartographic Association*, pp. 1–16, 2019. [Online]. Available: https://www.adv-cartogr-giscience-int-cartogr-assoc.net/1/16/2019/

[38] J. Osorio, "The Contagion of Drug Violence: Spatiotemporal Dynamics of the Mexican War on Drugs," *Journal of Conflict Resolution*, vol. 59, no. 8, pp. 1403–1432, 2015.

[39] V. D'Orazio and Y. Lin, "Forecasting conflict in africa with automated machine learning systems," *International Interactions*, vol. 48, no. 4, pp. 714–738, 2022. [Online]. Available: https://doi.org/10.1080/03050629.2022.2017290

[40] P. T. Brandt, V. D'Orazio, L. Khan, Y.-F. Li, J. Osorio, and M. Sianan, "Conflict forecasting with event data and spatio-temporal graph convolutional networks," *International Interactions*, vol. 48, no. 4, pp. 800–822, 2022. [Online]. Available: https://doi.org/10.1080/03050629.2022.2036987

[41] H. Mueller and C. Rauh, "Using past violence and current news to predict changes in violence," *International Interactions*, vol. 48, no. 4, pp. 579–596, 2022. [Online]. Available: https://doi.org/10.1080/03050629.2022.2063853

[42] C. Baden, C. Pipal, M. Schoonvelde, and M. A. C. G. van der Velden, "Three gaps in computational text analysis methods for social sciences: A research agenda," *Communication Methods and Measures*, vol. 16, no. 1, pp. 1–18, 2022. [Online]. Available: https://doi.org/10.1080/19312458.2021.2015574

[43] M. S. U. Miah, M. M. Kabir, T. B. Sarwar, M. Safran, S. Alfarhood, and M. Mridha, "A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and llm," *Scientific Reports*, vol. 14, no. 1, p. 9603, 2024.

[44] C. Liu, W. Zhang, Y. Zhao, A. T. Luu, and L. Bing, "Is translation all you need? a study on solving multilingual tasks with large language models," 2024. [Online]. Available: https://arxiv.org/abs/2403.10258

[45] M. Ziemski, M. Junczys-Dowmunt, and B. Pouliquen, *The United Nations Parallel Corpus v1.0*. Portorož, Slovenia: European Language Resources Association (ELRA), 2016. [Online]. Available: https://conferences.unite.un.org/uncorpus/Content/Doc/un.pdf

[46] M. Tkachenko, M. Malyuk, A. Holmanyuk, and N. Liubimov, "Label Studio: Data labeling software," 2020, open source software available from https://github.com/heartexlabs/label-studio. [Online]. Available: https://github.com/heartexlabs/label-studio

[47] Google cloud translation api. [Online]. Available: https://cloud.google.com/translate

[48] DeepL, "Deepl translator," https://www.deepl.com/translator. [Online]. Available: https://www.deepl.com/translator

[49] Deep Translator, "deep-translator: A flexible free and unlimited python tool to translate between different languages in a simple way using multiple translators," https://github.com/nidhaloff/deep-translator, 2020, gitHub. [Online]. Available: https://github.com/nidhaloff/deep-translator

[50] J. Tiedemann and S. Thottingal, "OPUS-MT – Building open translation services for the World: Annual Conference of the European Association for Machine Translation," *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 479–480, 2020, place: Geneva Publisher: European Association for Machine Translation. [Online]. Available: https://www.aclweb.org/anthology/2020.eamt-1.61/

[51] I. Caswell and B. Liang, "Recent advances in google translate," *Google Research Blog*, 2020.

[52] DeepL, "How does deepl work?" 2021.

[53] S. Swathi and L. Jayashree, "Machine translation using deep learning: A comparison," in *Proceedings of International Conference on Artificial Intelligence, Smart Grid and Smart City Applications (AISGSC)*. Springer, Cham, 2020, pp. 1–10. [Online]. Available: https://link.springer.com/article/10.1007/s10579-018-9410-z

[54] Google Cloud, "Google cloud translation api documentation," https://cloud.google.com/translate/docs, 2020, accessed: 2024-10-07.

[55] J. Tiedemann, "Parallel data, tools and interfacesin opus." in *Proceedings of LREC*, 2012.

[56] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040

[57] M. Post, "A call for clarity in reporting BLEU scores," *CoRR*, vol. abs/1804.08771, 2018. [Online]. Available: http://arxiv.org/abs/1804.08771

[58] A. Kim and J. Kim, "Vacillating human correlation of SacreBLEU in unprotected languages," in *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, A. Belz, M. Popović, E. Reiter, and A. Shimorina, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1–15. [Online]. Available: https://aclanthology.org/2022.humeval-1.1

[59] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, Eds. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72. [Online]. Available: https://aclanthology.org/W05-0909

[60] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with BERT," *CoRR*, vol. abs/1904.09675, 2019. [Online]. Available: http://arxiv.org/abs/1904.09675

[61] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," in *ICLR*, 2020. [Online]. Available: https://openreview.net/pdf?id=r1xMH1BtvB

[62] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, arXiv:1907.11692 [cs]. [Online]. Available: http://arxiv.org/abs/1907.11692

[63] S. Häffner, M. Hofer, M. Nagl, and J. Walterskirchen, "Introducing an Interpretable Deep Learning Approach to Domain-Specific Dictionary Creation: A Use Case for Conflict Prediction," *Political Analysis*, pp. 1–19, Mar. 2023, publisher: Cambridge University Press.

[64] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez, "Spanish pre-trained bert model and evaluation data," in *PML4DC at ICLR 2020*, 2020.

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

[66] L. Lingard, S. Cristancho, E. K. Hennel, C. St-Onge, and M. van Braak, "When english clashes with other languages: Insights and cautions from the writer's craft series," pp. 347–351. [Online]. Available: https://doi.org/10.1007/S40037-021-00689-2

[67] W. S. Cleveland, "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, vol. 74, no. 368, pp. 829–836, Dec. 1979, publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1979.10481038. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1979.10481038

[68] Microsoft, "Microsoft azure: Cloud computing services," 2024, accessed: 2024-10-10. [Online]. Available: https://azure.microsoft.com/

[69] Anthropic, "Claude 3 opus: A conversational ai model," 2023, accessed: 2024-10-10. [Online]. Available: https://www.anthropic.com/claude

[70] OpenAI, "Introducing chatgpt," https://openai.com/blog/chatgpt, 2022.

[71] I. Plaza, N. Melero, C. del Pozo, J. Conde, P. Reviriego, M. Mayor-Rocher, and M. Grandury, "Spanish and llm benchmarks: is mmlu lost in translation?" 2024. [Online]. Available: https://arxiv.org/abs/2406.17789

[72] T. J. Boerner, S. Deems, T. R. Furlani, S. L. Knuth, and J. Towns, "Access: Advancing innovation: Nsf's advanced cyberinfrastructure coordination ecosystem: Services & support," in *Practice and Experience in Advanced Research Computing*, ser. PEARC '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 173–176. [Online]. Available: https://doi.org/10.1145/3569951.3597559