IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 May 2026, accepted 22 June 2026, date of publication 10 July 2026, date of current version 24 July 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3712041

RESEARCH ARTICLE

Using English-Based NLP Tools for Domain-Specific Text in Foreign Languages

NAIF ALATRUSH[1], LUAY ABDELJABER[1], JAVIER OSORIO[2], DAGMAR HEINTZE[1], BRIKENA LIKO[2], PATRICK T. BRANDT[1], LATIFUR KHAN[1], (Fellow, IEEE), VITO J. D'ORAZIO[3], AND SULTAN ALSARRA[4]

[1]The University of Texas at Dallas, Richardson, TX 75080, USA
[2]University of Arizona, Tucson, AZ 85719, USA
[3]Political Science and Data Sciences, West Virginia University, Morgantown, WV 26506, USA
[4]Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 51178, Saudi Arabia

Corresponding authors: Javier Osorio (josorio1@arizona.edu) and Sultan Alsarra (salsarra@ksu.edu.sa)

This work was supported by U.S. National Science Foundation (NSF) Award under Grant 2311142; in part by the Delta at NCSA/University of Illinois through Allocation CIS220162 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services and Support (ACCESS) Program through NSF under Grant 2138259, Grant 2138286, Grant 2138307, Grant 2137603, and Grant 2138296; and in part by the Deanship of Scientific Research at King Saud University through the International Scientific Partnership Program (ISPP) Program under Grant ISPP25-16.

ABSTRACT Social scientists often machine-translate foreign-language texts into English and apply English-based natural language processing tools without systematically evaluating translation quality or annotation efficiency. To address this problem, this study provides evidence-based guidance for researchers applying English-centric natural language processing to domain-specific foreign-language corpora. We provide and empirically validate a structured framework combining multi-system machine translation evaluation and active learning for domain-specific text classification. Using 11,493 parallel Spanish and Arabic sentences aligned to English, we compare four machine translation systems (Google Translate, Deep, DeepL, OPUS) using SacreBLEU, METEOR, COMET, and BERTScore quality scores. Across languages and metrics, machine translation systems yield statistically comparable performance. We then evaluate eight active learning strategies using ConfliBERT for political conflict classification under a 20% annotation budget, corresponding to 1,155 samples from the training split. Binary classification exceeds F1 = 0.90, while QuadClass multi-class performance peaks around F1 $\approx$ 0.75. The Ensemble Intersection strategy achieves the highest performance in 53% of tasks and often matches or surpasses full-dataset results using only a fraction of labeled data. These results provide a practical workflow for researchers using English-based natural language processing tools on foreign-language, domain-specific corpora.

INDEX TERMS Active learning, guidelines, machine translation, quality metrics.

I. INTRODUCTION

Social scientists working with domain-specific texts in multiple languages often lack robust Natural Language Processing (NLP) tools for their native languages source texts. Beyond low-resource languages, scarcity of NLP tools is also a challenge for research requiring domain-specific NLP infrastructure that may not be available. To address this challenge, researchers often process their foreign language texts using machine translation (MT) into English and then use various NLP tools to analyze the MT English text. For example, scholars in political science using computational tools to analyze conflict processes require both NLP infrastructure in foreign languages and domain-specific NLP tools for analyzing conflict-related texts. Rather than using native language text to study foreign conflict, scholars often rely

The associate editor coordinating the review of this manuscript and approving it for publication was Li Zhang.

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

on MT text suitable for English-based NLP tools [1], [2]. In principle, this seems to be a cost-effective strategy taking advantage of rich English NLP tools. However, this approach seldom rests on a systematic assessment of the MT quality and often neglects translation errors.

This study provides guidelines and evaluation strategies to assist scholars applying English-based NLP tools to analyze domain-specific texts from other languages. The study makes three key contributions. *First*, it recommends different MT tools available to researchers. *Second*, it provides guidance on assessing the quality of the MT outputs using various quality evaluation scores. *Third*, since social scientists often apply their expertise through annotating the text, this research evaluates several Active Learning (AL) techniques to identify which maximizes performance while minimizing human annotation. These guidelines help researchers abandon ad-hoc and non-transparent data processing practices and encourage the adoption of systematic, transparent, evidence-based, computationally-backed strategies to inform the use of English-based NLP tools for domain-specific texts in non-English languages.

This research aims to systematically evaluate how English-based NLP tools can be applied to domain-specific texts written in foreign languages. Specifically, the research pursues three objectives: (1) to compare multiple MT tools when translating domain-specific texts into English; (2) to evaluate translation quality using several automatic machine translation quality metrics; and (3) to assess different active learning strategies to determine which approach maximizes classification performance while minimizing human annotation effort.

The work presented in this paper illustrates the use of these guidelines in political science using Spanish and Arabic texts about political conflict and violence, and leveraging English-language NLP tools to analyze them. The related works section discusses advances in MT tools, MT quality assessment, and AL. Then, the manuscript presents the data and discusses the downstream tasks. The next section discusses different MT tools and various quality metrics. The following part presents various AL strategies and evaluates their performance. The final section presents the conclusions.

II. RELATED WORK

A. RESEARCH ON MT AND MT EVALUATION

Most NLP research still centers on English, which limits access for lower-resource languages [3], [4], [5], [6], [7]. Scholars seldom systematically evaluate MT output, since human evaluation demands time and expertise and is hardly replicable [8].

Researchers have increasingly shifted from reference-based evaluation to reference-free MT quality estimation (QE). [9] trace QE evolution from handcrafted feature models to neural and Large Language Models (LLM). Reference-free methods such as COMET-QE [10] and TransQuest [11] illustrate this trend. MT evaluation has moved beyond human adequacy and fluency judgments [12], [13], [14] toward more scalable and consistent methods. For example, side-by-side comparative judgment such as SxS Multi-dimensional Quality Metrics improve inter-annotator agreement and error-marking consistency compared to standard point-wise evaluation [15]. To improve domain-specific MT, researchers have also explored domain adaptation strategies, such as fine-tuning on in-domain data, dynamic data selection, and terminology-constrained decoding [16], [17], [18].

Automatic MT evaluation has evolved from surface n-gram metrics, such as BLEU and METEOR, to embedding-based and neural approaches that capture semantic similarity more effectively, including BERTscore [19]. Special tests focus on anaphora and text coherence that standard metrics often miss [20]. More recent QE models such as COMET [21] and COMETKiwi [22], [23] predict adequacy and fluency without relying on reference translations. Multi-view fusion methods further strengthen reference-free QE by reducing bias and improving generalization across domains and languages [24]. Recent multilingual LLM approaches expand empirical baselines and data-mixing strategies for MT [25].

Recent studies show that standard MT quality scores often fail to predict LLM task performance and translation simplification can artificially boost results [26]. Reinforcement-learning such as COMETKiwi combine rule-based structural constraints with neural QE signals to guide reward optimization [27]. Additional work shows that domain-specific extractive LLMs outperform generic generative LLMs in political conflict classification tasks [28]. Researchers also use neural metrics beyond MT to assess human interlingual interpreting. Metrics such as BLEURT [29], COMET-22 [30], and xCOMET [31] correlate strongly with expert rater scores, confirming that they capture semantic fidelity [32].

Several studies compare human and automatic QE to analyze how closely automatic metrics align with expert judgments or fail to capture translation nuances [33], [34], [35], [36]. The Workshop on Machine Translation shared task further investigates the meta-evaluation process, revealing persistent spurious correlations and bias [37], [38]. Comparative studies examine different MT systems across languages and domains, showing that architecture, training data, and domain specificity all shape translation accuracy [26], [39], [40], [41], [42].

Introducing parameter-efficient adaptation methods such as sparse Mixture-of-Experts (MoE) improves multilingual translation while preventing catastrophic forgetting [43]. Recent instruction-tuned LLMs serve as strong translators, achieving domain-specific performance through prompt-based and minimal fine-tuning techniques [44]. Other developments include Large Reasoning Models integrating chain-of-thought reasoning for context- and intent-aware translation [45]. Hybrid designs such as LaMaTE pair LLM encoders with lightweight decoders to improve efficiency and generalization [46]. Adaptive few-shot prompting refines in-context translation by retrieving semantically similar

examples and reranking outputs for better low-resource MT performance [47]. Finally, compositional prompting decomposes complex sentences into shorter, independently translated segments before recombining them to boost MT quality [48].

Despite these impressive MT advances, significant gaps remain. Domain-specific studies show that MT reduces vocabulary richness and artificially boosts LLM performance by simplifying the text [26], [49]. However, most MT evaluation research still focuses on high-resource languages and general-domain corpora, leaving domain-specific effects largely untested. Moreover, even the most recent neural metrics are validated mainly through their correlation with human adequacy and fluency judgments, with little evidence that higher scores lead to better downstream performance in tasks such as political-event extraction or active-learning-based classification. Finally, the intersection of MT and AL itself remains largely unexplored, despite its potential to reduce annotation costs in low-resource, high-stakes settings. This study addresses these gaps by comparing multiple MT systems and evaluation metrics on UN political-conflict corpora and by analyzing how MT quality interacts with AL for cost efficient annotation in low-resource domains.

B. RELATED WORK ON ACTIVE LEARNING

Active Learning is widely used to reduce labeling costs by identifying and annotating the most informative samples [50]. Conventional AL techniques such as least confidence, margin sampling, and entropy use uncertainty-based acquisition functions to prioritize uncertain instances [51], [52]. These approaches are effective across multiple domains, but their reliance on a single acquisition heuristic motivated the development of more complex AL strategies [53].

Political text classification is highly challenging. Tasks such as event extraction, bias detection, or framing analysis involve nuanced language, implicit ideological cues, and context-dependent interpretations that vary across contexts [54], [55], [56]. Annotating political data is difficult and costly, and often requires expert coders. Although recent efforts use synthetic political data generation [57], questions remain about how such synthetic data faithfully captures the richness and nuance of real-world political discourse. To address the limitations of single-heuristic methods, [58] proposed ensemble AL strategies combining multiple acquisition functions—such as uncertainty, diversity, and representativeness—to leverage their complementary strengths. These ensemble methods provide more robust and effective recommendations for human annotation.

A recent study proposes a method called Active Semi-supervised Learning with a Debiasing mechanism (ASDE) [59]. This approach helps researchers choose the most useful examples to label first and then improves the model using both labeled and unlabeled data. It also includes steps to reduce bias that can spread during training. The goal is to reduce the amount of manual labeling required while still maintaining strong classification accuracy, especially in settings where labeled data are limited. Reference [60] report that dynamically selecting context-relevant few-shot examples and combining an ensemble of classification and generative models yields improved performance on subjective knowledge-grounded dialogue tasks, outperforming prior dialogue benchmarks.

III. TASKS AND DATA

This research relies on the United Nations Parallel Corpus (UNPC) [61] as curated by [49]. UNPC is a large collection of United Nations (UN) official documents between 1990 and 2014. It was produced and manually translated by UN professional translators. Most documents are available in all six official languages of the UN (English, Spanish, Arabic, French, Russian, and Chinese). The fully aligned UNPC corpus at the sentence-level contains 86,307 documents with a total of 11,365,709 sentences [61]. These cross-lingual UN documents are used as ''ground truth'' to assess the quality of different MT tools and AL strategies in the domain application. [49] annotated a randomly selected subset of UNPC text containing 11,493 parallel sentences in English, Spanish, and Arabic with the assistance of 12 human coders with political science expertise and bilingual skills in either English-Spanish or English-Arabic. Using the CAMEO ontology as conceptual framework [62], the coding process focused on classifying whether the sentences are relevant or not (binary classification), then classified relevant sentences within the QuadClass categorization of verbal/material-conflict/cooperation (multi-class classification), and finally disaggregated each QuadClass category into a mutually-exclusive binary classification. Appendix A provides additional data and annotation details.

As Figure 1a shows, the binary classification (BC) includes 52.4% relevant sentences and 47.6% non-relevant sentences. Coders annotated sentences according to QuadClass categories of verbal conflict, verbal cooperation, material conflict, or material cooperation based on the Conflict and Mediation Event Observation ontology [62]. While binary classification determines whether a sentence is relevant or non-relevant to political conflict, QuadClass annotations classify relevant sentences in a multi-class classification (MCC). Figure 1b shows that annotators classified 29.4% of the sentences as material conflict, 28.2% as material cooperation, 17.7% verbal conflict, and 24.8% verbal cooperation [49].

In a final annotation step, we refined the QuadClass annotations into BinQuad annotations, turning the MCC labels into a set of binary classifications taking the value of 1 for each QuadClass category and zero otherwise. This led to a set of simplified binary classifications for each of the verbal conflict (VERCONF), verbal cooperation (VERCOOP), material conflict (MATCONF), or material cooperation (MATCOOP) categories. These additional categories disaggregated the MCC task into specific categories for downstream research.

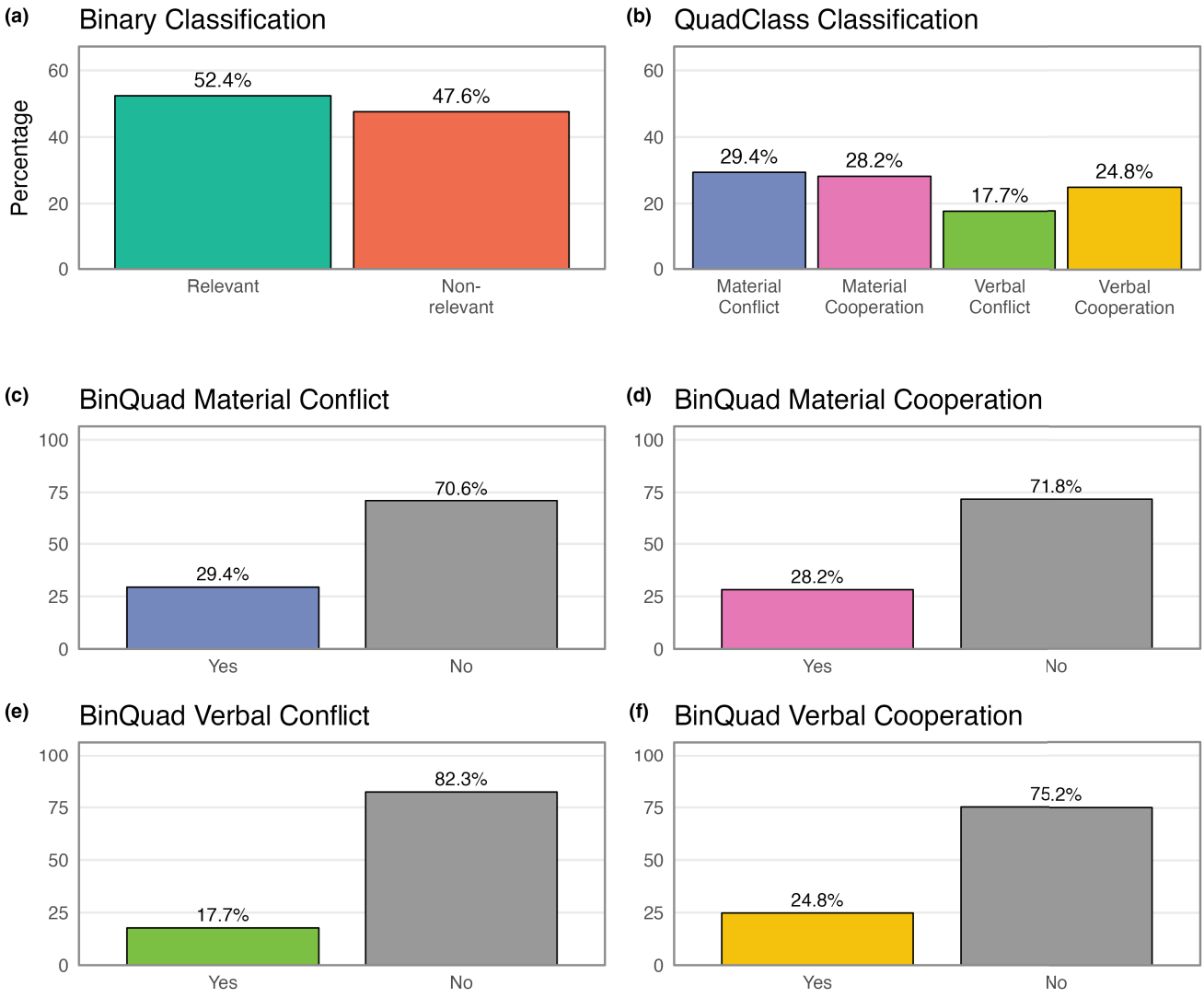


FIGURE 1. Classification Category Distributions.

The distribution of sentence annotations in Figure 1c has 29.4% as material conflict and 70.6% as not relevant for this category; Figure 1d shows 28.2% classified as material cooperation while 71.8% are non-relevant; Figure 1e indicates that 17.7% sentences are classified as verbal conflict while 82.3% are non-relevant; finally, Figure 1f reports 24.8% as verbal cooperation and the remainder 75.2% as non-relevant [49].

IV. MACHINE TRANSLATION

A. MACHINE-TRANSLATION TOOLS

Researchers working with non-English texts have multiple MT tools to process large amounts of text with high quality at relatively low cost. Conflict scholars studying political violence events and conflict processes often use machine translation to convert non-English reports and descriptions into English as input for computerized event coding [2], [62], [63], [64]. Social scientists often rely on a single MT tool without conducting a systematic evaluation of alternative MT tools and comparing their performance across non-English languages. Since different MT tools rely on different architectures and are informed by distinct training data, there may be questions about their capacity to preserve contextual meanings, grammatical and syntactic accuracy. If there are problems of translationese, machine-generated distortions, grammatical and syntactical changes, vocabulary reduction, or loss of domain-specific terms, then social science researchers need to evaluate the quality of the MT output [26], [65], [66], [67], [68].

Here are recommendations to assess systematically the quality of different MT tools on texts written in non-English languages.

1) **Select the non-English language text and identify the unit of analysis**. Here these are texts written in Arabic and Spanish in the UNPC database. The unit of analysis is at the sentence level.

2) **Use different MT tools to automatically translate the non-English text into English**. Rather than using a single MT tool, use an agnostic approach and compare multiple MT tools (detailed below).
3) **Use different MT quality assessment metrics** to systematically evaluate and compare the quality of the translated text. As discussed below, this study uses four translation quality metrics.

This study uses four MT methods to translate the Arabic and Spanish UNPC sentences into English. The first is Google Translate (GT) [69], a resource widely used by non-English speakers [70], [71], [72] and political scientists [2], [73], [74], [75], [76], [77]. GT uses Neural Machine Translation (NMT) on whole sentences, which improves the flow, contextual meaning, and accuracy of the translated output. The second is the DeepL API [78], a NMT-based tool regarded for its high accuracy and ability to capture contextual nuances because of its attention to the broader context, thus helping to disambiguate short phrases that lack sufficient context when considered in isolation. The third is the Deep Learning Translator (Deep), a Python library integrating various translation services by abstracting API complexities into a single unified architecture [79]. The fourth is OPUS [80], a Hugging Face library providing a large collection of state-of-the-art NMT models. In particular, we use the OPUS configuration with the `Helsinki-NLP/opus-mt-ar-en` and `Helsinki-NLP/opus-mt-es-en` models, which are specifically trained for Arabic to English and Spanish to English machine translation.

There are considerable variations in the design and functionality of the four different MT tools used in this study. Appendix B provides a detailed description of the technical characteristics of each MT tool explored here. These differences can in turn be consequential for the quality and reliability of their outputs.

The four MT tools evaluated all rely on transformer-based architectures, but have variations in their training corpora, network design, and user operational accessibility. These structural differences open the possibility of observing variations in the quality of the MT output generated by these different tools. The following section illustrates a practical application to systematically assess the quality of these different MT tools.

B. MACHINE-TRANSLATION QUALITY METRICS

Researchers can rely on human evaluation or computerized assessment tools, or a combination to assess the quality of the MT. Human evaluation allows careful assessments of the MT quality based on fluency, adequacy, syntactic and grammatical structure, and contextual accuracy [81]. Human raters can provide qualitative insights about systematic errors in grammatical and syntactical structures, cultural mismatches, or stylistic weaknesses that computerized tools may struggle to detect [82]. However, human quality assessment is highly costly and time consuming, involving professional translators, bilingual experts, or crowd-sourced evaluators, depending on the research design and available resources [83]. These high costs represent a considerable obstacle for scaling up the annotation process to large text collections.

As an alternative to human assessment, researchers can rely on various computerized tools to assess the quality of the MT output. Although not a substitute for human evaluation, computerized metrics provide quantitative indicators that offer comparability, objectivity, and scalability.

To assess the quality of the MT output, the four quality assessment scoring metrics are computed at the sentence level for each translation language, following [84]. The four are:

- **SacreBLEU:** [85], a variation of the BLEU (Bi-Lingual Evaluation Understudy) score, a pioneer MT quality metric [86] addressing tokenization and normalization challenges to ensure the evaluation is comparable and consistent across various systems [87]. This is the most rigid metric due to its reliance on the Moses tokenizer's heuristics and rules unique to a given language to normalize text and to handle punctuation or special characters.
- **METEOR:** (Metric for Evaluation of Translation with Explicit ORdering) [88], ordering the text to create a word alignment between the MT output and the native text, then calculating the similarity scores between pair sentences.
- **COMET:** (Crosslingual Optimized Metric for Evaluation of Translation) is a score of the MT semantic similarity based on pretrained LLMs [21]. This is a more flexible standard that prioritizes the similarity of word alignments and prioritizes meaning preservation.
- **BERTScore:** calculates the similarity of high-level semantic features between the MT output and native language text using BERT [89], favoring contextually correct translations and the use of synonyms without compromising correctness of the translation [19]. It is the most flexible MT quality assessment metric as it considers contextual correctness and synonyms as part of its evaluation.

Despite the differences among these metrics, they range from 0 to 1, with a higher value representing a greater similarity of the translated text to the ground truth or greater similarity between native and MT text [19], [90].

These quality assessment metrics use different architectures. SacreBLEU and METEOR are lexical-based metrics measuring the similarity between native text and the MT output based on mathematical or heuristic methods. COMET uses a neural network trained on human judgment to evaluate semantic accuracy and meaning preservation [21]. BERTScore uses an embedding-based metric measuring semantic similarity between native and MT text using the BERT architecture [91].

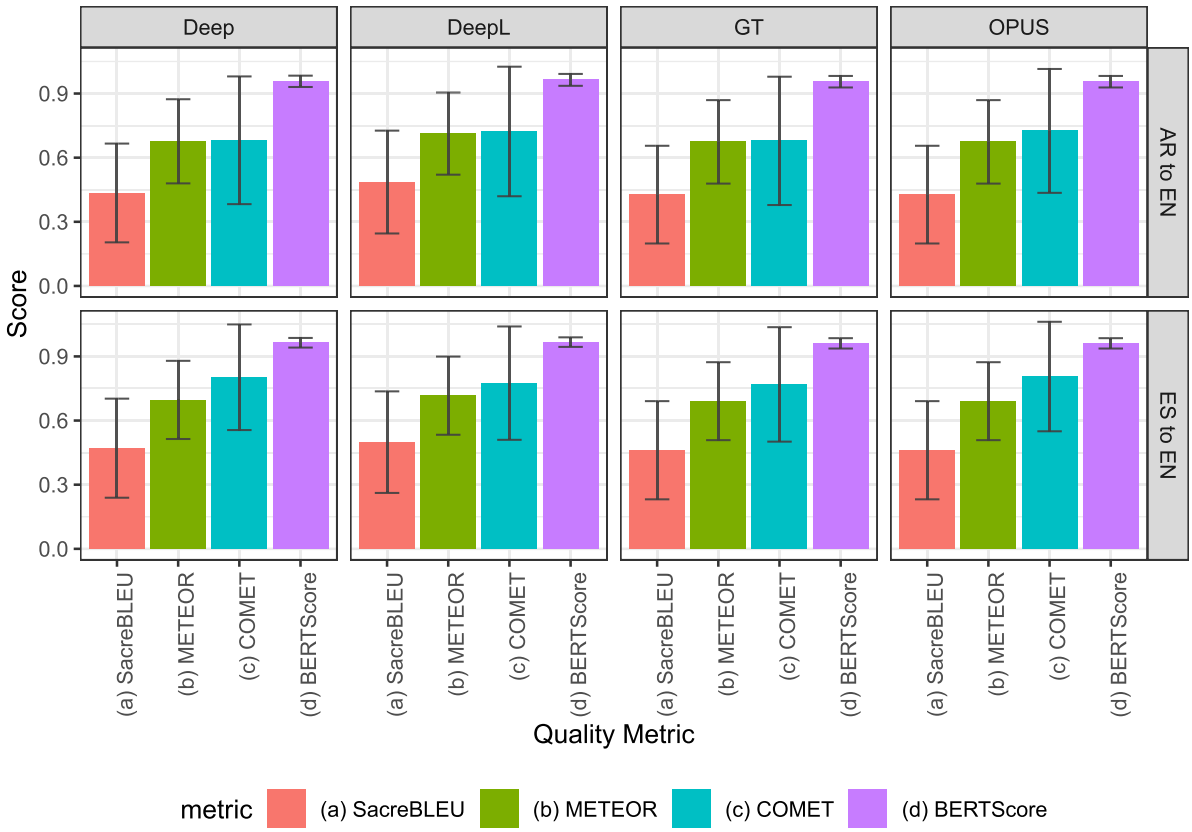


FIGURE 2. Machine Translation Quality Metrics.

C. QUALITY ASSESSMENT RESULTS

Figure 2 reports the quality translation scores for each translation language (from Arabic (AR) to English (EN) and from Spanish (ES) to English (EN)) on each MT tool (Deep, DeepL, GT, and OPUS) using each translation quality metric (SacreBLEU, METEOR, COMET, or BERTScore). The bars in Figure 2 indicate the average MT quality score across sentences with error bars for their standard deviation.

The metrics along the horizontal axis are ordered from the strictest (SacreBLEU) to the most flexible scale (BERTScore). SacreBLEU as expected consistently reports the lowest quality score across MT tools and languages, while BERTScore—the most flexible metric—has the highest translation quality.

Figure 2 shows that the different MT tools yield highly comparable translation outputs, particularly when considering the error bars for each quality score. Focusing on SacreBLEU (red bar) across MT tools in each language, results show that the output of different MT tools have similar means and all have overlapping standard errors.

Other MT tools report similar results, providing a valuable conclusion for researchers working with non-English texts since different MT tools yield highly similar results when processing text into English. Based on these quality metrics, researchers can rely on the MT tool most convenient or feasible to use, without systematically compromising the quality of the MT output. While this conclusion is only valid for the UNPC data in the selected languages and may be extensible to the domain of political conflict and cooperation, the generalizability of this claim should be empirically evaluated with other corpora and domains. These quantitative scores provide only an overview of translation quality without a qualitative analysis of translation distortions.

Based on Figure 2, it is possible to indicate that DeepL is marginally the best performing MT tool. For the remainder of the analysis, DeepL MT is used exclusively to illustrate the performance of different AL strategies.

V. ACTIVE LEARNING

A. ACTIVE LEARNING STRATEGIES

After selecting among MT tools and processing the entire corpus, researchers may proceed to identify and use different AL strategies to minimize their manual annotation costs while maximizing the performance of their effort. A pool-based AL framework begins with manually annotating a small initial dataset, which serves as the seed for model training. The remaining unlabeled instances form a large pool from which the model iteratively selects the most informative samples for annotation. At each iteration, the model evaluates all unlabeled samples and uses an acquisition function to

identify those that would most improve its performance. The selected samples are then manually annotated, added to the labeled dataset to retrain the model. This cycle continues until reaching the annotation budget. In this study, the total annotation budget is 1,155 training samples, representing about 20% of the training split after the 50/50 train-test partition. With 10 annotations per iteration, this results in 110 iterations for all strategies except Ensemble Union, which completes its selection process in 22 larger iterations due to its aggregated selection method.

All experiments use ConfliBERT [92] as the domain-specific NLP tool. ConfliBERT is a BERT-based model specifically designed for processing text related to political violence and conflict thus capturing domain-specific contextual nuances that are critical for defining political conflict events. Its specialized pretraining makes it particularly suitable for evaluating the interaction between MT quality and AL strategies in multilingual political text.

The AL framework aligns closely with the data distribution described above. Each task, such as detecting QuadClasses, relies on the natural distribution of categories in the dataset. Because these categories vary in frequency and linguistic clarity, the AL process adapts dynamically, prioritizing the most uncertain or representative samples for each label type. This ensures that the learned model remains both accurate and data efficient. We explore eight AL strategies:

- **Top Confidence Sampling:** selects samples with the lowest prediction confidence, assuming that uncertain predictions provide greater informational value for refining the model [51].
- **Maximum Entropy Sampling:** prioritizes instances with higher entropy, as they indicate greater ambiguity in classification and are expected to yield higher information gain when labeled [52].
- **Margin Sampling:** identifies samples based on the smallest difference between the top two predicted class probabilities. A narrow margin suggests uncertainty near decision boundaries, making such examples particularly valuable for improving classification performance [93].
- **Monte Carlo Dropout Sampling:** estimates predictive uncertainty by performing multiple stochastic forward passes with dropout enabled during inference. The variance across predictions serves as a proxy for model uncertainty, and samples with the highest variance are selected for labeling [94].
- **Core-set Sampling:** focuses solely on uncertainty, this method emphasizes diversity and representativeness. It selects samples that best cover the feature space, thereby ensuring that the labeled set is representative of the overall data distribution [95].
- **Bayesian Active Learning by Disagreement (BALD):** estimates the mutual information between model predictions and model parameters to quantify uncertainty. It selects instances that maximize disagreement across multiple predictions, striking a balance between uncertainty-driven exploration and diversity [96].
- **Ensemble Union:** combines the top-ranked candidates selected independently by multiple acquisition strategies. This method first selects the top m samples from each strategy, then removes duplicates from a larger candidate pool, then scores each candidate based on its relative rank, and makes the final selection from the aggregated pool. Ensemble Union accelerates exploration by incorporating diverse heuristics, promoting broader dataset coverage, and reducing the number of required AL rounds [58].
- **Ensemble Intersection:** prioritizes consensus by selecting only those samples jointly identified as highly informative by several other strategies. This approach yields a smaller but more reliable candidate set, as it focuses on instances with strong agreement across heuristics. Ensemble Intersection improves precision in sample selection and emphasizes robustness, making it particularly effective in later AL rounds [58].

B. ACTIVE LEARNING RESULTS

This section presents various F1 score statistics of the eight AL strategies across tasks. The discussion reports 1) *average F1* scores, 2) *maximum F1* score at any given sample size (providing the best possible result that researchers can get from different strategies) and 3) *graphical F1* for each AL strategy based on the sample size.

1) AVERAGE ACTIVE LEARNING PERFORMANCE

Table 1 reports the average F1 for each AL strategy by MT texts and tasks. For the Arabic to English MT using DeepL, results show that the Ensemble Intersection strategy reaches the best average score for the Binary Classification (BC) task. For the multi-class classification (MCC) of the different QuadClass categories, Max Entropy reports the best average performance. This strategy is also the best performer for the BinQuad tasks for Material Conflict (MATCONF) and Material Cooperation (MATCOOP) binary tasks. For the BinQuad tasks for Verbal Conflict (VERCONF) the Ensemble Intersection AL strategy has the highest average F1 score. Finally, Monte Carlo Dropout Sampling is the top performing AL strategy for Verbal Cooperation (VERCOOP). For the Spanish to English translation, the Ensemble Intersection AL strategy delivers the highest average F1 score for the relevant binary classification task. Results also show that the best approach for the QuadClass MCC is the BALD strategy. The Ensemble Intersection has the best average performance for the Material Conflict, Verbal Conflict, and Verbal Cooperation in the binary classification tasks. Max Entropy has the highest average results for the Material Cooperation in the binary task. Overall, the best performing AL strategies surpass the random benchmark and are very close to the full database results with only 20% of the data.

2) MAXIMUM ACTIVE LEARNING PERFORMANCE

Table 2 shows the maximum F1 score for each AL strategy per task in each MT language. For the Arabic to English

TABLE 1. Average F1 across tasks and AL strategies (values as percentages).

Dataset	Benchmark		Active Learning Strategies							
	Full Dataset	Random	Top Conf.	Max Ent.	MC Drop.	Margin	Coreset	BALD	Ens. Int.	Ens. Union
Arabic to English										
BC	91.73	89.16	90.81	90.73	90.29	91.00	90.12	89.89	**91.10**	91.02
MCC	76.13	73.58	73.64	**74.18**	73.93	74.13	73.27	73.99	73.56	71.79
MATCONF	92.91	89.67	91.42	**91.82**	91.56	91.40	91.18	91.32	91.72	91.21
MATCOOP	87.67	83.98	86.50	**86.99**	86.06	86.17	86.03	86.19	86.69	86.10
VERCONF	93.65	90.53	93.19	92.86	92.73	93.06	92.81	92.48	**93.20**	92.93
VERCOOP	90.30	87.73	89.65	89.19	**89.72**	89.65	89.48	88.92	89.50	88.98
Spanish to English										
BC	91.93	89.36	91.11	90.85	90.87	91.17	90.53	90.55	**91.38**	91.11
MCC	76.36	73.12	72.81	73.16	73.61	73.27	73.73	**73.76**	73.63	73.69
MATCONF	92.27	89.70	91.62	91.64	91.60	91.59	91.36	90.88	**91.68**	91.32
MATCOOP	87.90	84.89	86.54	**86.84**	85.91	86.57	85.68	85.78	86.14	86.05
VERCONF	93.44	91.02	92.74	92.83	92.71	92.93	92.57	91.30	**92.84**	92.45
VERCOOP	90.37	87.90	89.62	89.63	89.74	89.51	88.97	88.43	**89.80**	88.77

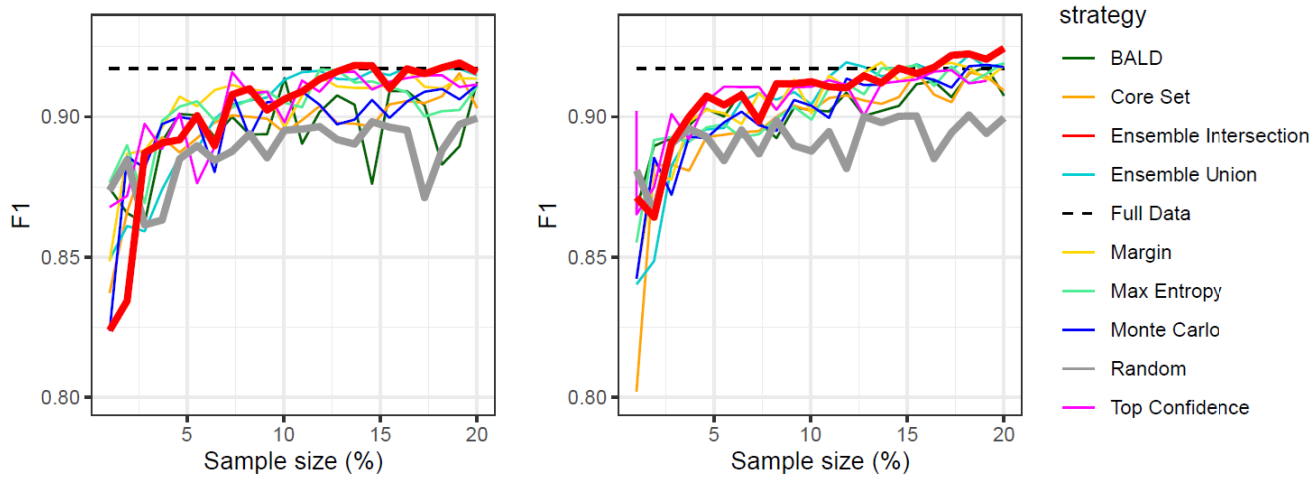
TABLE 2. Best F1 across datasets and AL strategies (values as percentages).

Dataset	Benchmark		Active Learning Strategies							
	Full Dataset	Random	Top Conf.	Max Ent.	MC Drop.	Margin	Coreset	BALD	Ens. Int.	Ens. Union
Arabic to English										
BC	91.73	89.95	91.61	91.68	91.14	91.68	91.55	91.38	**91.92**	91.82
MCC	76.13	74.28	75.32	75.80	75.58	75.52	74.80	75.29	**75.99**	75.81
MATCONF	92.91	89.99	92.03	**92.36**	92.35	92.20	91.98	92.08	92.28	92.08
MATCOOP	87.67	84.99	87.41	87.60	87.00	87.36	86.92	87.07	**87.61**	87.37
VERCONF	93.65	91.75	93.48	93.21	93.14	93.48	93.34	93.25	**93.49**	93.40
VERCOOP	90.30	88.84	**90.29**	89.89	90.21	90.09	90.26	89.46	90.27	90.10
Spanish to English										
BC	91.93	90.06	91.66	91.92	91.85	91.94	91.61	91.78	**92.44**	92.19
MCC	76.36	74.29	75.15	75.39	75.02	75.46	75.36	74.80	**75.61**	**75.61**
MATCONF	92.27	90.01	92.25	92.23	92.23	92.08	92.11	91.67	**92.29**	92.24
MATCOOP	87.90	85.40	87.21	**87.91**	86.67	87.60	86.84	86.50	87.24	87.40
VERCONF	93.44	91.54	93.26	93.22	93.07	93.23	93.31	92.01	**93.32**	93.26
VERCOOP	90.37	88.71	90.37	90.20	**90.37**	90.18	89.77	89.56	90.30	90.25

translation, results for the binary classification task show that Ensemble Intersection reaches the highest performance. With only 20% of the data, the Ensemble Intersection strategy surpasses the performance of using the full dataset. The Ensemble Intersection strategy also delivers the highest performance for the multi-class classification. For the Material Conflict BinQuad task, the Max Entropy strategy delivers the maximum F1 score. Ensemble Intersection again delivers the best results for the Material Cooperation and Verbal Conflict BinQuad. Finally, results for the Verbal Cooperation binary task indicate that Top Confidence Sampling shows the best performance. For Spanish to English text, the Max Entropy strategy yields the maximum F1 score. For the Material Cooperation binary classification task, Max Entropy also delivers the best results. Ensemble Intersection reports the top performance for the QuadClass MCC task, the Material Conflict binary task, and for Verbal Conflict. Finally, the best performing strategy for the binary Verbal Cooperation task is Monte Carlo Dropout sampling. In general, results for the Spanish to English MT show that AL strategies reach and often outperform the full data with only a fraction of the data.

3) FASTEST MAXIMUM

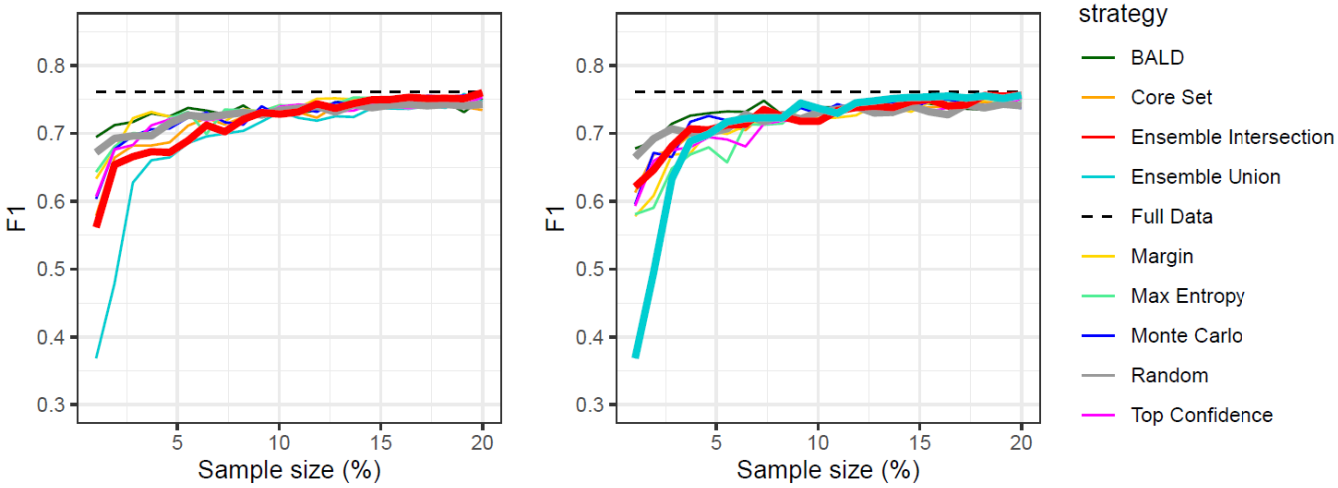
The data visualizations in Figures 3-6 show the variations in F1 performance by AL strategies against increases in the sample size up to 20% of the data. This identifies the best performing AL strategies at the lowest sample size in each task. This is what we call Fastest Maximum. As a baseline, all these plots have a horizontal black dashed line for performance based on the full data and a gray F1 score



(a) Binary Classification Arabic to English

(b) Binary Classification Spanish to English

FIGURE 3. AL for Binary Classification.



(a) Multi-Class Classification Arabic to English

(b) Multi-Class Classification Spanish to English

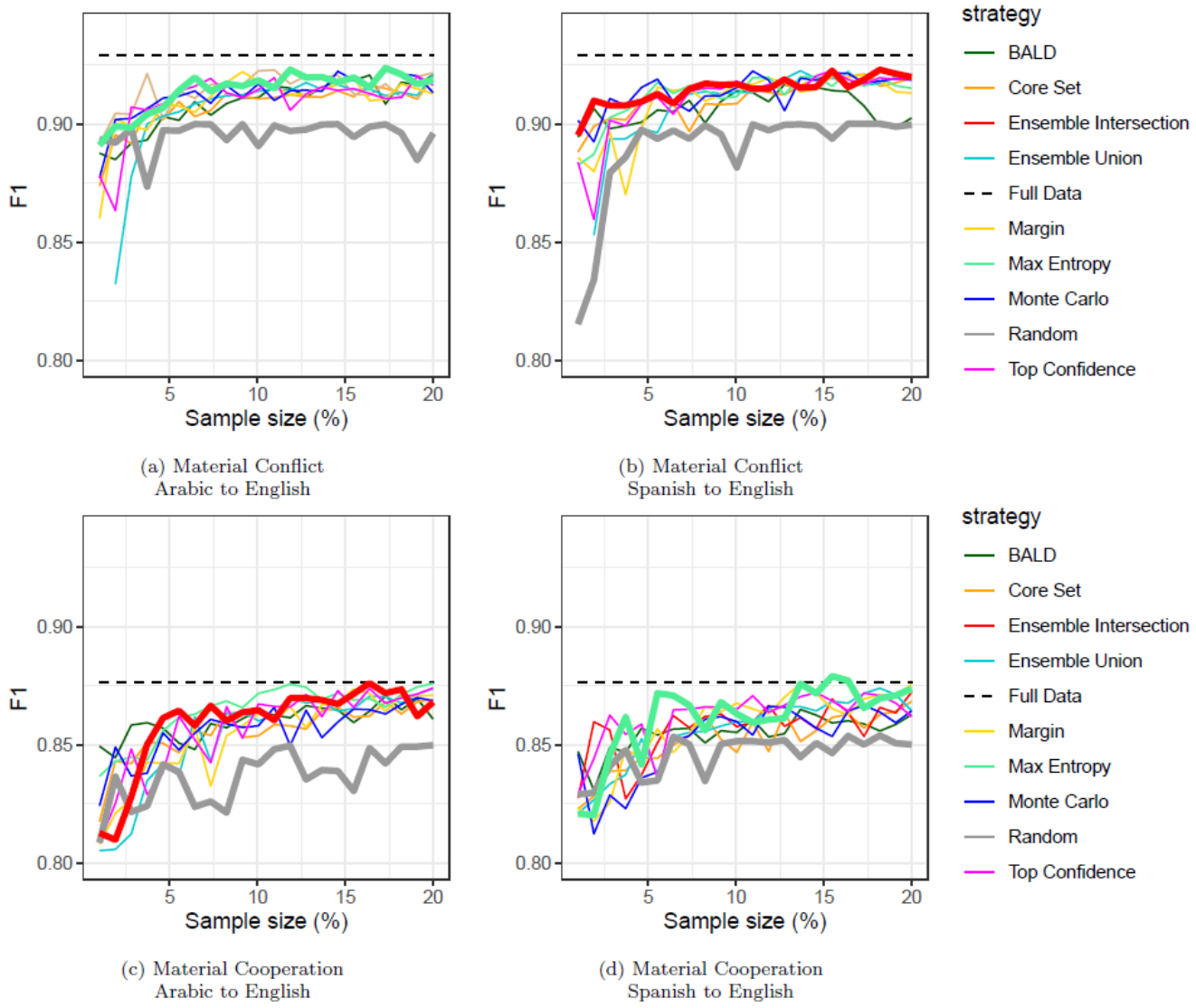
FIGURE 4. AL for Multi-Class Classification.

for random selection. The other colored lines are for the 8 proposed AL strategies at different percentages of the sample. In each plot, the best performing AL strategy is highlighted with its corresponding color.

Figure 3 presents the different AL results for binary classification, showing that this is an easy task with an F1 score generally higher than 0.90. Panel 3a shows that the Ensemble Intersection is the top performing AL strategy for the Arabic to English MT, reaching an F1 score of 0.9192 at 19.1% of the sample. Panel 3b also indicates that Ensemble Intersection yields the best performance for MT text from Spanish to English, reaching an F1 of 0.9244 at 20% of the sample.

Figure 4 reports the F1s for the QuadClass MCC task across sample sizes. Results suggest this is a harder task since all different AL strategies top out with an F1 around 0.75, below those of the binary classification task. This is an important insight for researchers since the increasing difficulty of the task may be associated with lower AL performance. For Arabic to English translation Figure 4a shows that Ensemble Intersection has the top performance at the smallest sample size relative to the baselines. Figure 4b shows that both the Ensemble Intersection and Union strategies achieve top AL performance at the same sample size.

Figure 5 shows the F1 across AL sample size for the binary classification of Material Conflict and Material Cooperation for Arabic and Spanish MT into English. The higher performance of the top two panels in Figure 5 compared to the relatively lower performance in the lower panels suggests



(a) Material Conflict Arabic to English

(b) Material Conflict Spanish to English

(c) Material Cooperation Arabic to English

(d) Material Cooperation Spanish to English

FIGURE 5. AL for Material Conflict and Cooperation.

that classifying Material Conflict sentences in either Arabic or Spanish translations into English yields relatively better results than classifying Material Cooperation tasks. Figure 5a shows that Max Entropy provides the best results for the Arabic to English translation for Material Conflict. Results for this same task in the Spanish to English translation in Figure 5b indicate that Ensemble Intersection is the top performer. For Material Cooperation in the Arabic to English text, Figure 5c also shows that Ensemble Intersection yields the best performance for annotation purposes. Finally, Figure 5d indicates that Max Entropy achieves the best results with Material Cooperation from Spanish to English.

Figure 6 presents the results for the BinQuad Verbal Conflict and Cooperation tasks. The top two panels in Figure 6 show relatively better performance than the bottom panels, suggesting that classifying MT Verbal Conflict tasks is relatively easier than classifying Verbal Cooperation ones. Figure 6a indicates that the Top Confidence strategy has the better performance for the Verbal Conflict task in the Arabic to English MT data. Figure 6b indicates that the Ensemble Intersection AL strategy yields the best performance for this same classification task in the Spanish to English text. In the Verbal Cooperation BinQuad task, Figure 6c indicates that Top Confidence is the best AL tool for this task in the Arabic to English text. Finally, Figure 6d shows that Monte Carlo Dropout yields the best results for classifying Verbal Cooperation in text translated from Spanish to English.

Table 3 summarizes the performance of the AL strategies across tasks and MT choices. The Ensemble Intersection strategy yields the best performance across tasks, translated languages, and metrics. The Ensemble Intersection AL approach stands as the top performing strategy in 53% of

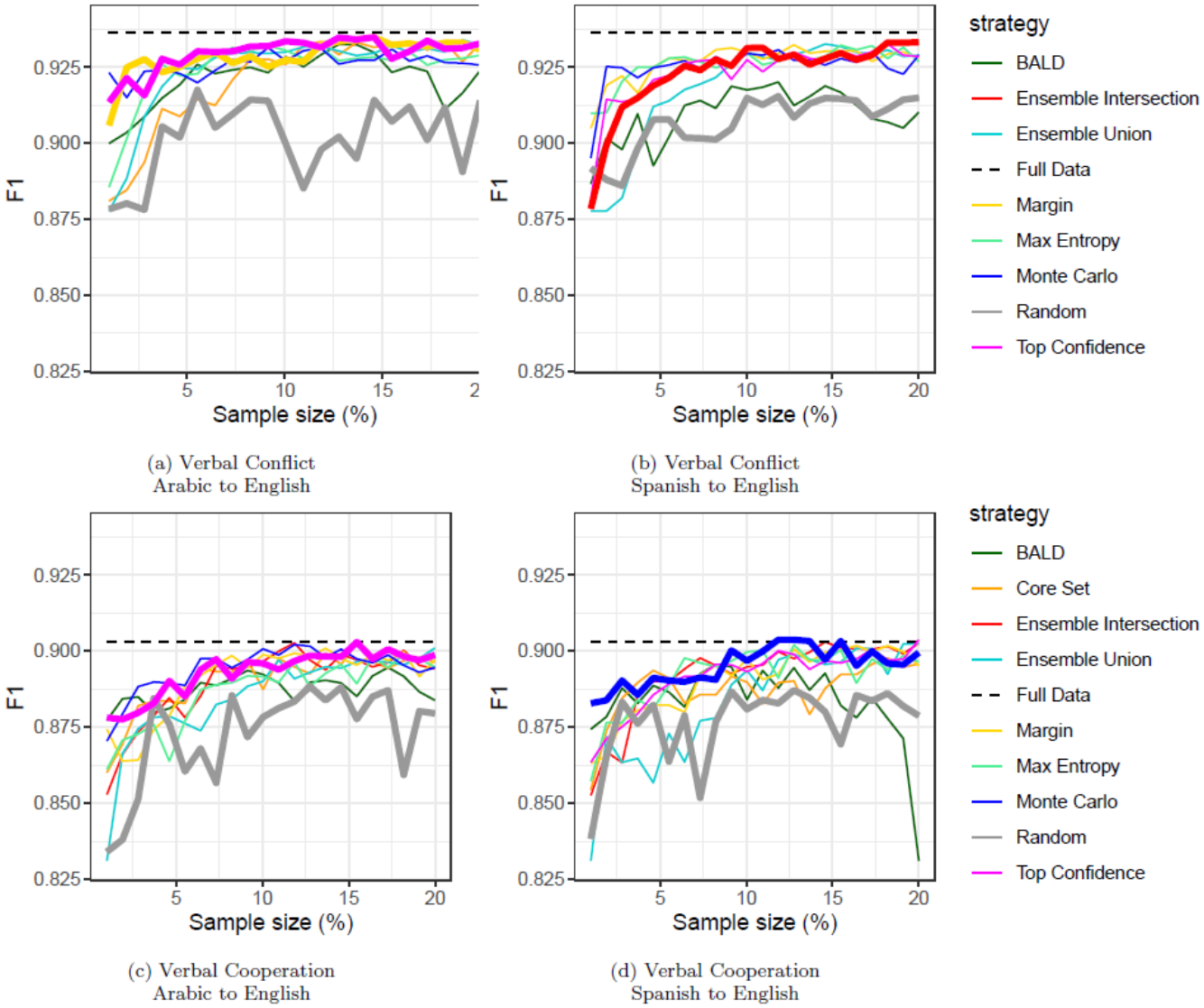


FIGURE 6. AL for Verbal Conflict and Cooperation.

the experiments. This is more than twice the rate of Max Entropy, the second top performing strategy delivering the best results only in 25% of the experiments. This assessment offers valuable guidelines for researchers working with native texts in low-resource languages. It suggests the use of an Ensemble Intersection strategy to maximize classification scores while minimizing the manual annotations required.

VI. IMPLICATIONS

The guidelines advanced in this study provide three central contributions to promote more rigorous and transparent workflows. First, it recommends that researchers adopt an agnostic approach by considering multiple MT tools. Since MT performance may vary across languages or domains, researchers would benefit from evaluating various MT tools for their subject of interest. Second, it encourages social science researchers to apply multiple MT assessment metrics to evaluate the quality of each MT output. Comparing the quality of different MT texts across metrics offers a systematic and evidence-based approach to select the best MT tool for a given language and domain. Third, researchers should evaluate the quality of different AL strategies to assist their annotation efforts. Evaluating these strategies helps identify those generating the best performance while minimizing annotation costs. This study shows that the Ensemble Intersection is the most effective AL strategy for the texts and classification tasks analyzed in this study.

VII. CONCLUSION

Social scientists frequently encounter a major obstacle when working with foreign-language texts: the lack of reliable, high-quality NLP infrastructure in those languages. This

TABLE 3. AL results summary.

	Average	Maximum	Fastest maximum
Arabic to English			
BC	Ensemble Intersection	Ensemble Intersection	Ensemble Intersection
MCC	Max Entropy	Ensemble Intersection	Ensemble Intersection
MATCONF	Max Entropy	Max Entropy	Max Entropy
MATCOOP	Max Entropy	Ensemble Intersection	Ensemble Intersection
VERCONF	Ensemble Intersection	Ensemble Intersection	Top Confidence
VERCOOP	Monte Carlo Dropout	Top Confidence	Top Confidence
Spanish to English			
BC	Ensemble Intersection	Ensemble Intersection	Ensemble Intersection
MCC	BALD	Ensemble Intersection	Ensemble Intersection
MATCONF	Ensemble Intersection	Ensemble Intersection	Ensemble Intersection
MATCOOP	Max Entropy	Max Entropy	Max Entropy
VERCONF	Ensemble Intersection	Ensemble Intersection	Ensemble Intersection
VERCOOP	Ensemble Intersection	Monte Carlo Dropout	Monte Carlo Dropout

challenge becomes even more pronounced in topics that rely on specialized input materials and require domain-specific NLP tools. To address this limitation, researchers often translate non-English texts into English using MT tools and then apply the wide array of English-language NLP tools to analyze the translated outputs. However, despite the popularity of this workaround, many researchers adopt it in an ad hoc manner, without articulating clear methodological justifications for their choices. These opaque methodological decisions can substantially affect both the quality of the research and the validity of its findings.

This paper responds to these challenges by offering a structured set of guidelines and evaluation methods designed to help social scientists who depend on English-centric NLP for analyzing non-English or domain-specific texts. It argues that MT should be treated as a critical modeling choice rather than a routine or arbitrary pre-processing step. Researchers are therefore encouraged to systematically and transparently document and justify their MT tools, parameter configurations, quality evaluation metrics, and manual annotation procedures to ensure methodological rigor and reproducibility.

To illustrate the application of these guidelines, we analyze a collection of domain-specific texts in Spanish and Arabic about political conflict and violence from the UN Parallel Corpus. We use different MT tools, quality assessment scores, and AL strategies on these parallel texts using English-centric NLP tools. Results indicate that no single MT tool yields the best performance. These findings suggest that the translation tools examined in this study (GT, Deep, DeepL, and OPUS) produce comparably accurate translations from Spanish and Arabic into English within this domain. The consistent results across MT quality evaluation metrics (SacreBLEU, METEOR, COMET, and BERTScore) reinforce this conclusion. Finally, to help researchers maximize efficiency, we compare eight AL strategies across languages and classification tasks, finding that the Ensemble Intersection strategy most consistently maximizes classification performance while minimizing annotation effort.

TABLE 4. Annotation tasks according to the CAMEO framework.

		Cooperation	Conflict
Relevant	**Verbal**	Participants issued a joint press release	She accused the Prime Minister
	Material	Police activated their coordination protocol	Rebels ambushed military forces
Not relevant	The farmer collected his belongings		

Overall, these guidelines encourage researchers using English-centric NLP to analyze domain-specific non-English texts to embrace an agnostic approach that considers multiple tools and strategies at each stage of the research process, calls for transparent reporting of technical specifications and performance results, and emphasizes evidence-based justifications for all methodological decisions.

DATA AVAILABILITY

The United Nations Parallel Corpus (UNPC) is publicly available [61].

APPENDIX A
DATA AND ANNOTATION PROCESS

The empirical foundations of this study rely on the data curated and annotated by [49] based on a large collection of documents from the United Nations Parallel Corpus (UNPC) [61]. UNPC contains official documents issued by the United Nations (UN) between 1990 and 2014 with parallel sentences in all six official UN languages (English, Spanish, Arabic, French, Russian, and Chinese) professionally translated by professional interpreters and translators of

TABLE 5. Example of Different MT Outputs.

	Spanish	Arabic
UNPC in ES & AR	En Colombia, Chile, el Paraguay y el Perú también se han promulgado leyes que otorgan facultades especiales para la investigación a las fuerzas del orden.	وقد سنّت باراغواي وبيرو وشيلي وكولومبيا تشريعات تمنح سلطات خاصة في مجال التحقيق للسلطات المعنية بإنفاذ القانون
MT using GT	Laws granting special investigative powers to law enforcement agencies have also been enacted in Colombia, Chile, Paraguay and Peru.	Paraguay, Peru, Chile and Colombia have enacted legislation granting special investigative powers to law enforcement authorities.
MT using DeepL	Colombia, Chile, Paraguay and Peru have also enacted laws granting special investigative powers to law enforcement.	Chile, Colombia, Paraguay, Peru, Chile, and Colombia have enacted legislation granting special investigative powers to law enforcement authorities.
MT using Deep	Laws granting special investigative powers to law enforcement agencies have also been enacted in Colombia, Chile, Paraguay and Peru.	Paraguay, Peru, Chile and Colombia have enacted legislation granting special investigative powers to law enforcement authorities.
MT using OPUS	In Colombia, Chile, Paraguay and Peru, laws have also been enacted granting special investigative powers to law enforcement officials.	Chile, Colombia, Paraguay and Peru have enacted legislation granting special investigative powers to law enforcement authorities.
UNPC in EN	Colombia, Chile, Paraguay and Peru have also enacted legislation giving their law enforcement authorities special investigative powers.	

the UN Department for General Assembly and Conference Management (DGACM) Translation Services. In total, the UNPC collection comprises 86,307 documents disaggregated into 11,365,709 sentences fully aligned across languages. To curate this data, [49] selected a random sample of 11,493 UNPC sentences from resolutions issued by the United Nations Security Council related to human rights, the protection of civilians, and terrorism, three topics highly relevant to political conflict and violence. For the purpose of this study, we use only the subset of sentences originally produced in Spanish, Arabic, and English.

The manual annotation process relied on a team of 12 human coders with domain expertise in political science and international relations, and bilingual skills in either English-Spanish or English-Arabic, who processed the 11,493 UNPC sentences following a strict annotation and validation protocol. To conduct the annotations, human coders used Label Studio [97], an annotation interface capable of processing text in multiple languages. Coders were grouped in triplets to simultaneously code sequential batches of 300 sentences randomly selected. Each coder annotated all the sentences in a batch independently from the other two coders who blindly coded the same sentences. After all coders completed all the blind annotations in a batch, the manager revealed the annotations across coders to identify overlap and discrepancies. This allowed coders to compare their annotation choices. Then, coders discussed and resolved discrepancies until reaching 100% agreement. The best performing coder in each triad made the final decision on a small number of complex cases or unresolved discrepancies. The overall labeling process reported a 92% inter-coder reliability score, thus providing confidence about the high quality of the annotation output.

Based on the CAMEO ontology [62], the industry standard for classifying international political interactions, the labeling process consisted of three annotation tasks:

1) Classify whether a sentence is relevant or not (binary classification) with respect to political interactions as conceptualized in the CAMEO ontology.
2) For relevant sentences, classify what type of QuadClass category best captures the content of the text (multi-class classification). The QuadClass indicates if an international political interaction is verbal cooperation, material cooperation, verbal conflict, or material conflict.
3) Each QuadClass is then disaggregated into a set of mutually exclusive binary classifications referred here as BinQuads, indicating whether an event corresponds to verbal cooperation or not, material cooperation or not, verbal conflict or not, or material conflict or not.

Table 4 illustrates the conceptual framework with sample sentences for the relevant and non-relevant classification along with the four QuadClass categories of relevant sentences. Based on the examples provided in Table 4, any of the sentences included in the Relevant area would be effectively labeled as relevant, while the residual sentences at the bottom of the Table would be classified as not relevant. Then, for any of the relevant sentences, Table 4 provides examples of the types of sentences that would correspond to any of the QuadClass categories of verbal cooperation, verbal conflict, material cooperation, and material conflict.

The resulting annotation process generated a database containing binary classification for relevance, multi-class classification for the QuadClass categories, and a set of binary classification variables for each of the BinQuad category. Refer to Figure 1 in the body of the manuscript for a data visualization of the different annotations across tasks.

APPENDIX B
MACHINE TRANSLATION TOOLS

As discussed in the body of the manuscript, this study relies on four machine translation tools including Google Translate (GT) [69], DeepL API [78], Deep Learning Translator (Deep) [79], and OPUS [80].

For GT, the methodological strategy relies on the subscription-based Google Cloud Translation service. Its system architecture integrates transformer-based neural networks with a decoder Recurrent Neural Network (RNN), thereby combining strengths of both paradigms. Another GT feature is its continual retraining based on massive text corpora gathered through embedding-based web crawls [98], [99], [100]. This data collection process employs a miner optimized for precision over recall, ensuring that the model prioritizes accurate and context-sensitive translations even if some breadth of coverage is sacrificed [101].

For DeepL, we used the subscription-based Graphical User Interface (GUI). DeepL operates on artificial neural networks partly based on transformers with some architectural innovations. This architecture is combined with training on carefully curated datasets and supplementary web-crawled corpora to enable DeepL deliver translations that are often more fluent and contextually accurate than those produced by other MT tools [102]. Furthermore, DeepL incorporates internal quality-control mechanisms, including automated filtering and human-in-the-loop evaluation, which strengthen its grammatical and syntactical consistency, and overall meaning reliability.

The Deep Learning Translator is a Python library that integrates multiple translation services. Rather than relying on a single system, it connects several translators including Google Translate, DeepL, Microsoft Translator, Yandex, Baidu, Linguee, Papago, and others. The package also supports diverse input formats, provides batch translation functionalities, and leverages API servers for processing large datasets [79]. For our analysis, we specifically used the Google Translate option. However, unlike subscription-based Google Cloud services, this free translator is vulnerable to throttling issues and temporary breakdowns, as documented in prior evaluations [103], [104].

Finally, the OPUS MT system relies on a transformer-based NMT trained on the openly available OPUS parallel corpora, a massive collection of 1,212 corpora comprising 1,004 languages and about 59 billion sentence pairs [105]. In contrast to proprietary MT tools that do not completely disclose their training corpus, the OPUS parallel corpora repository consists of freely accessible bilingual text pairs collected from a wide range of sources, making the models highly transparent and reproducible. While OPUS' reliance on open data is a positive feature that promotes accessibility and adaptability, the varying quality of some of its corpora can introduce challenges in ensuring consistent translation performance across domains.

Table 5 presents an example of a single UNPC sentence originally written in Spanish (center column) and Arabic (right column) and their subsequent translations using GT, DeepL, Deep, and OPUS, and finally showing the original sentence in English as included in the UNPC corpus. As the Table illustrates, each MT tool generates slight variations in the translation output derived from the original source languages in Spanish and Arabic. While the output of each of the MT tool resembles the meaning of the same sentence originally written in English, each MT output has slight syntactical and grammatical variations.

ACKNOWLEDGMENT

This material is based upon High Performance Computing (HPC) resources supported by the University of Arizona TRIF, UITS, and Research, Innovation, and Impact (RII). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF, NCSA, or the affiliated university resource providers.

REFERENCES

[1] E. Boschee, J. Lautenschlager, S. O'Brien, S. Shellman, and J. Starz, "ICEWS weekly event data," Harvard Dataverse, Cambridge, MA, USA, 2018, doi: 10.7910/DVN/QI2T9A.

[2] A. Halterman, B. E. Bagozzi, A. Beger, P. Schrodt, and G. Scarborough, "Plover and polecat: A new political event ontology and dataset," SocArXiv, Center Open Sci., Charlottesville, VA, USA, Tech. Rep., Apr. 2023, doi: 10.31235/osf.io/rm5dw.

[3] A. Søgaard, "Should we ban English NLP for a year?" in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Dec. 2022, pp. 5254–5260.

[4] G. I. Winata, A. F. Aji, S. Cahyawijaya, R. Mahendra, F. Koto, A. Romadhony, K. Kurniawan, D. Moeljadi, R. E. Prasojo, P. Fung, T. Baldwin, J. H. Lau, R. Sennrich, and S. Ruder, "NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages," in *Proc. 17th Conf. Eur. Chapter Assoc. Comput. Linguistics*, May 2023, pp. 815–834.

[5] F. Guzmán, P.-J. Chen, M. Ott, J. Pino, G. Lample, P. Koehn, V. Chaudhary, and M. Ranzato, "The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Nov. 2019, pp. 6098–6111.

[6] H. J. Haynie, D. Blasi, H. Skirgård, S. J. Greenhill, Q. D. Atkinson, and R. D. Gray, "Grambank's typological advances support computational research on diverse languages," in *Proc. 5th Workshop Res. Comput. Linguistic Typology Multilingual NLP*, Croatia, May 2023, pp. 147–149.

[7] S. M. Ul Qumar, M. Azim, S. M. K. Quadri, M. Alkanan, M. S. Mir, and Y. Gulzar, "Deep neural architectures for Kashmiri-English machine translation," *Sci. Rep.*, vol. 15, no. 1, pp. 21–30014, Aug. 2025.

[8] A. Mukherjee and M. Shrivastava, "Lost in translation? Found in evaluation: A comprehensive survey on sentence-level translation evaluation," *ACM Comput. Surv.*, vol. 58, no. 1, pp. 1–47, Jan. 2026.

[9] H. Zhao, Y. Liu, S. Tao, W. Meng, Y. Chen, X. Geng, C. Su, M. Zhang, and H. Yang, "From handcrafted features to LLMs: A brief survey for machine translation quality estimation," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2024, pp. 1–10.

[10] E. A. Chimoto and B. A. Bassett, "COMET-QE and active learning for low-resource machine translation," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, Dec. 2022, pp. 4735–4740.

[11] T. Ranasinghe, C. Orasan, and R. Mitkov, "TransQuest: Translation quality estimation with cross-lingual transformers," in *Proc. 28th Int. Conf. Comput. Linguistics*, Spain, Dec. 2020, pp. 5070–5081.

[12] M. Martindale and M. Carpuat, "Fluency over adequacy: A pilot study in measuring user trust in imperfect MT," in *Proc. 13th Conf. Assoc. Mach. Transl. Amer.*, Mar. 2018, pp. 13–25.

[13] M. Muftah, "Machine vs human translation: A new reality or a threat to professional Arabic–English translators," *PSU Res. Rev.*, vol. 8, no. 2, pp. 484–497, Nov. 2024.

[14] S. Seljan, I. Dunder, and M. Pavlovski, "Human quality evaluation of machine-translated poetry," in *Proc. 43rd Int. Conv. Inf., Commun. Electron. Technol. (MIPRO)*, Sep. 2020, pp. 1040–1045.

[15] Y. Song, P. Riley, D. Deutsch, and M. Freitag, "Enhancing human evaluation in machine translation with comparative judgement," in *Proc. 63rd Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2025, pp. 20536–20551.

[16] C. Chu and R. Wang, "A survey of domain adaptation for neural machine translation," in *Proc. 27th Int. Conf. Comput. Linguistics*, Aug. 2018, pp. 1304–1319.

[17] P. Koehn and R. Knowles, "Six challenges for neural machine translation," in *Proc. 1st Workshop Neural Mach. Transl.*, Aug. 2017, pp. 28–39.

[18] C. Chu, R. Dabre, and S. Kurohashi, "An empirical comparison of domain adaptation methods for neural machine translation," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2017, pp. 385–391.

[19] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," 2020, *arXiv:1904.09675*.

[20] R. Bawden, R. Sennrich, A. Birch, and B. Haddow, "Evaluating discourse phenomena in neural machine translation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, Jun. 2018, pp. 1304–1313.

[21] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "COMET: A neural framework for MT evaluation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Nov. 2020, pp. 2685–2702.

[22] R. Rei, M. Treviso, N. M. Guerreiro, C. Zerva, A. C. Farinha, C. Maroti, J. G. C. de Souza, T. Glushkova, D. Alves, L. Coheur, A. Lavie, and A. F. T. Martins, "CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task," in *Proc. 7th Conf. Mach. Transl. (WMT)*, Dec. 2022, pp. 634–645.

[23] R. Rei, N. M. Guerreiro, J. Pombal, D. van Stigt, M. Treviso, L. Coheur, J. G. C. de Souza, and A. Martins, "Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task," in *Proc. 8th Conf. Mach. Transl.*, Dec. 2023, pp. 841–848.

[24] H. Huang, S. Wu, K. Chen, X. Liang, H. Di, M. Yang, and T. Zhao, "Multi-view fusion for universal translation quality estimation," *Inf. Fusion*, vol. 102, Feb. 2024, Art. no. 102022.

[25] M. Cui, P. Gao, W. Liu, J. Luan, and B. Wang, "Multilingual machine translation with open large language models at practical scale: An empirical study," in *Proc. Conf. Nations Americas Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, Apr. 2025, pp. 5420–5443.

[26] J. Osorio, A. Alshammari, N. Alatrush, D. Heintze, A. Converse, S. Alsarra, L. Khan, P. T. Brandt, and V. D'Orazio, "The devil is in the details: Assessing the effects of machine-translation on llm performance in domain-specific texts," in *Proc. 20th Mach. Transl. Summit*, vol. 1, 2025, pp. 315–332.

[27] Z. Feng, S. Cao, J. Ren, J. Su, R. Chen, Y. Zhang, Z. Xu, Y. Hu, J. Wu, and Z. Liu, "MT-R1-zero: Advancing LLM-based machine translation via R1-zero-like reinforcement learning," 2025, *arXiv:2504.10160*.

[28] P. T. Brandt, S. Alsarra, V. D'Orazio, D. Heintze, L. Khan, S. Meher, J. Osorio, and M. Sianan, "Extractive versus generative language models for political conflict text classification," *Political Anal.*, vol. 34, no. 3, pp. 344–372, Jul. 2026.

[29] T. Sellam, D. Das, and A. Parikh, "BLEURT: Learning robust metrics for text generation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2020, pp. 7881–7892.

[30] R. Rei, J. G. C. de Souza, D. Alves, C. Zerva, A. C. Farinha, T. Glushkova, A. Lavie, L. Coheur, and A. F. T. Martins, "COMET-22: Unbabel-IST 2022 submission for the metrics shared task," in *Proc. 7th Conf. Mach. Transl. (WMT)*, Dec. 2022, pp. 578–585.

[31] D. Larionov, M. Seleznyov, V. Viskov, A. Panchenko, and S. Eger, "XCOMET-lite: Bridging the gap between efficiency and quality in learned MT evaluation metrics," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Nov. 2024, pp. 21934–21949.

[32] C. Han and X. Lu, "Beyond BLEU: Repurposing neural-based metrics to assess interlingual interpreting in tertiary-level language learning settings," *Res. Methods Appl. Linguistics*, vol. 4, no. 1, Apr. 2025, Art. no. 100184.

[33] P. Koehn and C. Monz, "Manual and automatic evaluation of machine translation between European languages," in *Proc. Workshop Stat. Mach. Transl.-StatMT*, Jun. 2006, pp. 102–121.

[34] M. Przybocki, K. Peterson, and S. Bronsart, "2008 NIST metrics for machine translation (METRICSMATR08) development data," Linguistic Data Consortium, Philadelphia, PA, USA, Tech. Rep. LDC2009T05, 2009.

[35] T. Kocmi, C. Federmann, R. Grundkiewicz, M. Junczys-Dowmunt, H. Matsushita, and A. Menezes, "To ship or not to ship: An extensive evaluation of automatic metrics for machine translation," in *Proc. 6th Conf. Mach. Transl.*, Nov. 2021, pp. 478–494.

[36] D. Deutsch, J. Juraska, M. Finkelstein, and M. Freitag, "Training and meta-evaluating machine translation evaluation metrics at the paragraph level," in *Proc. 8th Conf. Mach. Transl.*, Singapore, Dec. 2023, pp. 996–1013.

[37] S. Perrella, L. Proietti, A. Scirè, E. Barba, and R. Navigli, "Guardians of the machine translation meta-evaluation: Sentinel metrics fall in!" in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics*, Aug. 2024, pp. 16216–16244.

[38] N. Moghe, A. Fazla, C. Amrhein, T. Kocmi, M. Steedman, A. Birch, R. Sennrich, and L. Guillou, "Machine translation meta evaluation through translation accuracy challenge sets," *Comput. Linguistics*, vol. 51, no. 1, pp. 73–137, Mar. 2025.

[39] A. Toral, F. Gaspari, S. K. Naskar, and A. Way, "A comparative evaluation of research vs. online MT systems," in *Proc. 15th Annu. Conf. Eur. Assoc. Mach. Transl.*, May 2011, pp. 13–20.

[40] Y. Bansal, B. Ghorbani, A. Garg, B. Zhang, M. Krikun, C. Cherry, B. Neyshabur, and O. Firat, "Data scaling laws in NMT: The effect of noise and architecture," 2022, *arXiv:2202.01994*.

[41] S. Sato, J. Sakuma, N. Yoshinaga, M. Toyoda, and M. Kitsuregawa, "Vocabulary adaptation for domain adaptation in neural machine translation," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, Nov. 2020, pp. 4269–4279.

[42] S. Alsarra, M. Alrashoud, J. Osorio, V. D'Orazio, L. Khan, and P. T. Brandt, "Multilingual approaches to extractive question answering in political texts," *Comput. Math. Org. Theory*, vol. 31, no. 4, pp. 299–322, Dec. 2025.

[43] S. Zhu, L. Pan, D. Jian, and D. Xiong, "Overcoming language barriers via machine translation with sparse mixture-of-experts fusion of large language models," *Inf. Process. Manage.*, vol. 62, no. 3, May 2025, Art. no. 104078.

[44] D. Vilar, M. Freitag, C. Cherry, J. Luo, V. Ratnakar, and G. Foster, "Prompting PaLM for translation: Assessing strategies and performance," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2023, pp. 15406–15427.

[45] S. Liu, C. Lyu, M. Wu, L. Wang, W. Luo, K. Zhang, and Z. Shang, "New trends for modern machine translation with large reasoning models," 2025, *arXiv:2503.10351*.

[46] Y. Luo, T. Zheng, Y. Mu, B. Li, Q. Zhang, Y. Gao, Z. Xu, P. Feng, X. Liu, T. Xiao, and J. Zhu, "Beyond decoder-only: Large language models can be good encoders for machine translation," 2025, *arXiv:2503.06594*.

[47] L. Tang, J. Qin, W. Ye, H. Tan, and Z. Yang, "Adaptive few-shot prompting for machine translation with pre-trained language models," 2025, *arXiv:2501.01679*.

[48] A. Zebaze, B. Sagot, and R. Bawden, "Compositional translation: A novel LLM-based approach for low-resource machine translation," 2025, *arXiv:2503.04554*.

[49] J. Osorio, S. Alsarra, A. Converse, A. Alshammari, D. Heintze, L. Khan, N. Alatrush, P. T. Brandt, V. D'Orazio, N. Zawad, and M. Billah, "Keep it local: Comparing domain-specific LLMs in native and machine translated text using parallel corpora on political conflict," in *Proc. 2nd Int. Conf. Found. Large Lang. Models (FLLM)*, Nov. 2024, pp. 542–552.

[50] L. Abdeljaber, N. Alatrush, S. Alsarra, M. Billah, A. Alshammari, and L. Khan, "Active learning for medical text classification: Insights from ophthalmology and broader medical domains," in *Proc. IEEE 13th Int. Conf. Healthcare Informat. (ICHI)*, Jun. 2025, pp. 216–227.

[51] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Proc. SIGIR*, 1994, pp. 3–12.

[52] B. Settles, "Active learning literature survey," Dept. Comput. Sci., Univ. of Wisconsin, Madison, WI, USA, Tech. Rep. 1648, Jan. 2009.

[53] A. Shelmanov, D. Puzyrev, L. Kupriyanova, D. Belyakov, D. Larionov, N. Khromov, O. Kozlova, E. Artemova, D. V. Dylov, and A. Panchenko, "Active learning for sequence tagging with deep pre-trained models and Bayesian uncertainty estimates," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics*, Apr. 2021, pp. 1698–1712.

[54] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, Mar. 2018.

[55] P. T. Brandt and M. Sianan, "Measurement of event data from text," *Frontiers Political Sci.*, vol. 6, Jan. 2025, Art. no. 1453640.

[56] F. Hamborg, K. Donnay, and B. Gipp, "Automated identification of media bias in news articles: An interdisciplinary literature review," *Int. J. Digit. Libraries*, vol. 20, no. 4, pp. 403–429, Dec. 2019.

[57] A. Halterman, "Synthetically generated text for supervised text analysis," *Political Anal.*, vol. 33, no. 3, pp. 181–194, Jul. 2025.

[58] N. Alatrush, S. Alsarra, A. Alshammari, L. Abdeljaber, N. Zawad, L. Khan, P. T. Brandt, J. Osorio, and V. J. D'Orazio, "Advancing active learning with ensemble strategies," in *Proc. Recent Adv. Natural Lang. Process.*, 2025, pp. 57–66.

[59] Y. Chen, T. Zhou, D. Zeng, S. Li, K. Liu, and J. Zhao, "ASDE: Low-budget text classification via active semi-supervised learning with debiasing training mechanism," *Inf. Process. Manage.*, vol. 63, no. 2, Mar. 2026, Art. no. 104390.

[60] D. Rao, J. Zhuang, and Z. Jiang, "Leveraging dynamic few-shot prompting and ensemble method for task-oriented dialogue with subjective knowledge," *Inf. Process. Manage.*, vol. 63, no. 2, Mar. 2026, Art. no. 104317.

[61] M. Ziemski, M. Junczys-Dowmunt, and B. Pouliquen, "The united nations parallel corpus v1.0," Eur. Lang. Resour. Assoc. (ELRA), Paris, France, Tech. Rep., 2016.

[62] 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," in *Proc. Int. Stud. Assoc.*, New Orleans, LA, USA, 2002.

[63] E. Boschee, J. Lautenschlager, S. O'Brien, S. Shellman, J. Starz, and M. Ward, "ICEWS coded event data," Harvard Dataverse, Cambridge, MA, USA, 2016, doi: 10.7910/DVN/28075.

[64] S. P. O'Brien, "Crisis early warning and decision support: Contemporary approaches and thoughts on future research," *Int. Stud. Rev.*, vol. 12, no. 1, pp. 87–104, Mar. 2010.

[65] E. Vanmassenhove, D. Shterionov, and A. Way, "Lost in translation: Loss and decay of linguistic richness in machine translation," in *Proc. 17th Mach. Transl. Summit, Res. Track*, Aug. 2019, pp. 222–232.

[66] E. Vanmassenhove, D. Shterionov, and M. Gwilliam, "Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics*, Apr. 2021, pp. 2203–2213.

[67] J. Luo, C. Cherry, and G. Foster, "To diverge or not to diverge: A morphosyntactic perspective on machine translation vs human translation," *Trans. Assoc. Comput. Linguistics*, vol. 12, pp. 355–371, Apr. 2024.

[68] M. Artetxe, G. Labaka, and E. Agirre, "Translation artifacts in cross-lingual transfer learning," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Nov. 2020, pp. 7674–7684.

[69] (2025). *Google Cloud Translation API*. [Online]. Available: https://cloud.google.com/translate

[70] Y. Wu et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*.

[71] B. Klimova, "Use of machine translation in foreign language education," *Cogent Arts Humanities*, vol. 12, no. 1, Dec. 2025, Art. no. 2491183.

[72] C. Rossi and J.-P. Chevrot, "Uses and perceptions of machine translation at the European commission," *J. Specialised Transl.*, no. 31, pp. 177–200, Jan. 2019.

[73] C. Lucas, R. A. Nielsen, M. E. Roberts, B. M. Stewart, A. Storer, and D. Tingley, "Computer-assisted text analysis for comparative politics," *Political Anal.*, vol. 23, no. 2, pp. 254–277, 2015.

[74] Z. M. Almahasees, "Assessing the translation of Google and Microsoft bing in translating political texts from arabic into English," *Int. J. Lang., Literature Linguistics*, vol. 3, no. 1, pp. 97–100, Mar. 2017.

[75] E. de Vries, M. Schoonvelde, and G. Schumacher, "No longer lost in translation: Evidence that Google translate works for comparative bag-of-words text applications," *Political Anal.*, vol. 26, no. 4, pp. 417–430, Oct. 2018.

[76] D. Maier, C. Baden, D. Stoltenberg, M. De Vries-Kedem, and A. Waldherr, "Machine translation Vs. multilingual dictionaries assessing two strategies for the topic modeling of multilingual text collections," *Commun. Methods Measures*, vol. 16, no. 1, pp. 19–38, Jan. 2022.

[77] J. I. Plenter, "Advantages and pitfalls of machine translation for party research: The translation of party manifestos of European parties using DeepL," *Frontiers Political Sci.*, vol. 5, Nov. 2023, Art. no. 1268320.

[78] (2024). *DeepL Translator*. [Online]. Available: https://www.deepl.com/translator

[79] (2020). *Deep-Translator: A Flexible Free and Unlimited Python Tool to Translate Between Different Languages in a Simple Way Using Multiple Translators*. [Online]. Available: https://github.com/nidhaloff/deep-translator

[80] J. Tiedemann and S. Thottingal, "OPUS-MT–Building open translation services for the world: Annual conference of the European association for machine translation," in *Proc. 22nd Annu. Conf. Eur. Assoc. Mach. Transl.*, 2020, pp. 479–480.

[81] P. Koehn, *Statistical Machine Translation*. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[82] S. Castilho, M. Popović, and A. Way, "On context span needed for machine translation evaluation," in *Proc. 12th Lang. Resour. Eval. Conf.*, May 2020, pp. 3735–3742.

[83] Y. Graham, T. Baldwin, A. Moffat, and J. Zobel, "Continuous measurement scales in human evaluation of machine translation," in *Proc. 7th Linguistic Annotation Workshop Interoperability Discourse*, 2013, pp. 33–41.

[84] L. Han, G. Erofeev, I. Sorokina, S. Gladkoff, and G. Nenadic, "Examining large pre-trained language models for machine translation: What you Don't know about it," 2022, *arXiv:2209.07417*.

[85] M. Post, "A call for clarity in reporting BLEU scores," 2018, *arXiv:1804.08771*.

[86] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2002, pp. 311–318.

[87] A. Kim and J. Kim, "Vacillating human correlation of SacreBLEU in unprotected languages," in *Proc. 2nd Workshop Human Eval. NLP Syst. (HumEval)*, May 2022, pp. 1–15.

[88] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, Jun. 2005, pp. 65–72.

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

[90] E. Chatzikoumi, "How to evaluate machine translation: A review of automated and human metrics," *Natural Lang. Eng.*, vol. 26, no. 2, pp. 137–161, Mar. 2020.

[91] S. Lee, J. Lee, H. Moon, C. Park, J. Seo, S. Eo, S. Koo, and H. Lim, "A survey on evaluation metrics for machine translation," *Mathematics*, vol. 11, no. 4, p. 1006, Feb. 2023.

[92] Y. Hu, M. Hosseini, E. Skorupa Parolin, J. Osorio, L. Khan, P. Brandt, and V. D'Orazio, "ConfliBERT: A pre-trained language model for political conflict and violence," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2022, pp. 5469–5482.

[93] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden Markov models for information extraction," in *Proc. Int. Symp. Intell. Data Anal.*, 2001, pp. 309–318.

[94] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.

[95] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–13.

[96] N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel, "Bayesian active learning for classification and preference learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011, pp. 1–9.

[97] M. Tkachenko, M. Malyuk, A. Holmanyuk, and N. Liubimov. (2020). *Label Studio: Data Labeling Software*. [Online]. Available: https://github.com/heartexlabs/label-studio

[98] M. R. Costa-Jussá et al., "Scaling neural machine translation to 200 languages," *Nature*, vol. 630, pp. 841–846, Jun. 2024.

[99] B. Thompson, M. Dhaliwal, P. Frisch, T. Domhan, and M. Federico, "A shocking amount of the web is machine translated: Insights from multi-way parallelism," in *Proc. Findings Assoc. Comput. Linguistics ACL*, 2024, pp. 1763–1775.

[100] R. Van Noord, M. Esplá-Gomis, M. Chichirău, G. Ramírez-Sánchez, and A. Toral, "Quality beyond a glance: Revealing large quality differences between web-crawled parallel corpora," in *Proc. 31st Int. Conf. Comput. Linguistics*, 2025, pp. 1824–1838.
[101] I. Caswell and B. Liang, "Recent advances in Google translate," Google Res., Google LLC, Mountain View, CA, USA, Tech. Rep., Jun. 2020.
[102] *How Does Deepl Work?*, DeepL Team, Cologne, Germany, Oct. 2021.
[103] S. Swathi and L. S. Jayashree, "Machine translation using deep learning: A comparison," in *Proc. Int. Conf. Artif. Intell., Smart Grid Smart City Appl.*, 2020, pp. 389–395.
[104] (2026). *Cloud Translation API | Google Cloud Documentation*. [Online]. Available: https://docs.cloud.google.com/translate/docs/reference/rest
[105] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," in *Proc. LREC*, 2012, pp. 2214–2218.

NAIF ALATRUSH received the B.S. degree in network engineering and operating systems from Damascus University, Damascus, Syria, in 2008, the M.S. degree in management information systems from Arab Academy, Damascus, Syria, in 2012, and the M.S. degree in computer science from Oklahoma City University, Oklahoma City, OK, USA, in 2015. He is currently pursuing the Ph.D. degree in software engineering with The University of Texas at Dallas, Richardson, TX, USA. He has been working as a Software Engineer and Team Lead in the software industry, since 2015, with experience in enterprise software systems, intelligent automation, cloud integrations, and communication platforms. His research interests include natural language processing, large language models, active learning, machine learning, mixture-of-experts architectures, cybersecurity, and big data analytics. His recent work focuses on LLM-assisted active learning frameworks, ensemble learning strategies, and multilingual political text analysis. He is the author and co-author of multiple peer-reviewed publications in IEEE journals and international AI and NLP conferences. He received the Best Paper Award at SBP-BRiMS 2024 and was also a recipient of the Professors Award in Computer Science and the High Honor Award from Oklahoma City University, in 2015.

LUAY ABDELJABER received the Ph.D. degree in computer science from The University of Texas at Dallas, specializing in machine learning and artificial intelligence. He is a Researcher at the Ophthalmology Department, Stanford University School of Medicine, where he develops artificial intelligence and machine learning methods for healthcare applications. His research focuses on medical AI, natural language processing, clinical large language models, retrieval-augmented generation (RAG), and machine learning for healthcare data analysis. His work spans interdisciplinary areas including ophthalmology, dementia research, vaccine confidence modeling, and trustworthy AI systems for clinical applications. He has led and contributed to projects involving semi-supervised learning for ophthalmic image analysis, active learning for medical text classification, clinical NLP systems for extracting information from electronic health records, and large language model–based methods for healthcare research. His current interests include building clinically reliable AI systems that incorporate temporal reasoning, structured evidence extraction, and large-scale clinical data. He has authored and co-authored publications across healthcare AI, machine learning, multilingual natural language processing, political language modeling, and blockchain security. His research contributions include domain-specific language models, choroidal tumor classification, and retrieval-augmented approaches for analyzing vaccine misinformation and public health narratives.

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

DAGMAR HEINTZE received the Diploma degree in interpreting for English and Spanish from Johannes Gutenberg-Universität, Mainz, Germany, and the master's degree in European studies from Freie Universität Berlin, Germany. She is currently pursuing the Ph.D. degree in political science at The University of Texas at Dallas. From 2017 to 2022, she taught European Law and English for Law Enforcement as a tenured Lecturer at Brandenburg State Police University, and is currently on academic leave. Since 2024, she has been a Research Assistant at the multidisciplinary ConfliBERT Research Group with Dr. Patrick T. Brandt, The University of Texas at Dallas. Her publications span political science journals and computer science conference proceedings. Her research interests include computational methods, human rights, conflict, international law, and international organizations.

BRIKENA LIKO received the M.S. degree in linguistics from the University of Tirana, the M.S. degree in computational linguistics from Montclair State University, and the Ph.D. degree in generative syntax from the University of Tirana. She is currently pursuing the Ph.D. degree in computational linguistics and the M.S. degree in computer science with the University of Arizona. She was a Fulbright Student at Montclair State University. She is a Computational Linguist at the Academy of Sciences of Albania, where she works on national projects funded by the Government of Albania, including the Albanian Encyclopedia and the Contemporary Dictionary of Albanian. She is also a Research Assistant at the University of Arizona. She leads the design, development, and maintenance of the online platforms for these national language resources.

PATRICK T. BRANDT received the A.B. degree in government from the College of William and Mary, the M.S. degree in mathematical methods in the social sciences from Northwestern University, and the Ph.D. degree in political science from Indiana University. He is a Professor of political science at The University of Texas at Dallas. His expertise spans time series analysis, Bayesian statistics, and machine learning methods applied to international relations, political economy, conflict, terrorism, and public opinion. His research develops new statistical models to study change in political and economic events over time, including vector autoregression and count time series approaches. His work has been supported by the National Science Foundation and the Center for Risk and Economic Analysis of Terrorist Events (CREATE). He has published in leading journals of political science and methodology and serves on editorial boards, including *Journal of Conflict Resolution* and *International Interactions*. He was a recipient of multiple recognitions, including the Robert H. Durr Award from the Midwest Political Science Association and the 2024 UT Dallas Recognition of Outstanding Achievements in Research (ROAR) Award. In addition to research, he teaches and mentors students in advanced political methodology, and has led training workshops in Bayesian time series and forecasting models for the social sciences.

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

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

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