ConfliBERT

A Language Model for Political Conflict

Patrick T. Brandt[1] Sultan Alsarra[2] Vito J. D'Orazio[3] Dagmar Heintze[1]
Latifur Khan[4] Shreyas Meher[1] Javier Osorio[5] Marcus Sianan[1]

[1]UT Dallas, Economic, Political, and Policy Sciences

[2]King Saud University, Software Engineering

[3]West Virginia University, Political Science

[4]UT Dallas, Computer Science

[5]University of Arizona, Political Science

1. What is the problem we want to address
   - Key ideas
2. ConfliBERT as a coder
3. Example / Intro to the Method
4. GTD Event Classifications
   - Setup
   - Performance Evaluation
   - Results
   - Efficiency results
5. Conclusions
6. So you want to do this?

Opening Acknowledgments

Receipts:

This research was supported by NSF award 2311142 and 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 awards 2138259, 2138286, 2138307, 2137603, and 2138296. This material used 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 National Science Foundation, NCSA, or the affiliated university resource providers.

Political Science version

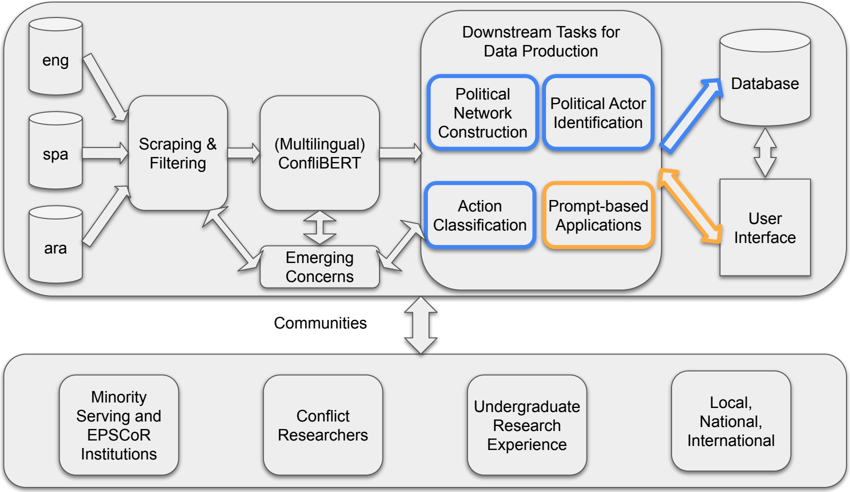
So you want to use AI / LLMs for your political science research?

There are two ways to look at the options:

Extractive LLMs / AI: This organizes or finds information in a set of sources (images, texts, etc.)

Generative LLM / AI: summarize and present conclusions or reasoning about sources.

Our Framework



Key ideas

Classification: Which texts contain relevant information about politics, conflict, violence? Two kinds

1. binary classifications: yes / no questions See Example
2. multi-label classifications: in a series of reports about protests, which types of protest are present (labor, peaceful, violent, etc.)?

Named Entity Recognition (NER): What are the "who" and "whom" that characterize the event? See Example

Masking / Coding new entities and / or events: extension of any ontology of new kinds of events.

LLM Options and Tasks

Extractive LLMs

- BERT (Google)
- RoBERTa
- DeBERT
- ... and many *BERTs

Generative LLMs

- ChatGPT (OpenAI)
- Claude (Anthropic)
- Llama (Meta), Gemma (Google) & Qwen (Alibaba)

Access via: Cloud APIs or Hugging Face

Backend: Ollama/llama.cpp (Generative) and Hugging Face (ConfliBERT)

Political ↔ Computer Science Versions

Questions to be answered

1. Is a thing present or discussed in a text, report, story, document?
   Binary Classification
2. Who / what / where is participating in a political event or discussed in a document?
   Named Entity Recognition
3. What or which attribute does an event, actor, or action have?
   Masking/Coding

DEFINED

ConfliBERT is a LLM trained on a *curated corpus of high-quality text data about politics, conflict, and violent events*

- Training examples / data link
- Testing examples / data link

Uses domain knowledge with which it has been "trained" to be a more useful language model than vanilla BERT, LLM, or a simple dictionary approach.

What is BERT

Ask the decoder what an encoder is (via Gemini)?

AI Overview

A BERT model, which stands for "Bidirectional Encoder Representations from Transformers," is a deep learning model developed by Google that excels at natural language processing (NLP) tasks by understanding the context of words within a sentence by analyzing both the words before and after it, allowing for a more nuanced interpretation of language compared to traditional methods; essentially, it learns to "read" text like a human does, considering the surrounding context to grasp the meaning of words.

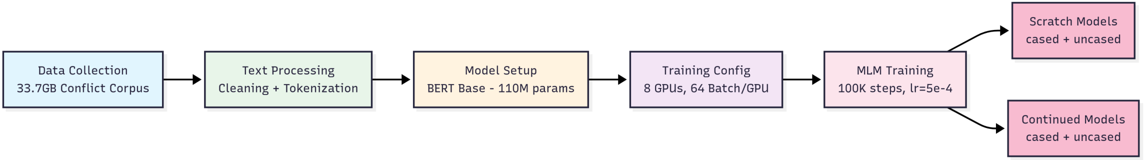
Key points about BERT:

Bidirectional processing:

Unlike older models that only looked at words before a target word, BERT analyzes both the preceding and following words to understand context fully.

Transformer architecture:

ConfliBERT LLM development chart



Try ConfliBERT!

Scan to Start

Basic Statement of the Problem

Task: You want to code (*terrorist attacks*) from some reports or texts.

- You have the texts in digital form with some meta data for collating them (source, time period, geography). So you have already done the *source selection*.
- You need to *extract information* relative to some ontology, codebook, or rules.
- It's too much or too *expensive* (in time, money, or iterative processing) to do it (again, more, etc.)

Coding data like GTD with an LLM

- Global Terrorism Dataset (GTD) is an appropriate application:
  - Comprehensive open-source database of terrorist events.
  - Contains example information for classification (what kind of an attack is in the event?)
  - Text is consistent and well structured
  - Text was expert classified (Codebook includes 'Who', 'what' and 'to whom')
- Nature of the data is suitable for NER and MC but not BC tasks.
- NER and BC from GTD text descriptions.
- Compares of model performance (ConfliBERT, LLama 3.1, Gemma 2, Qwen 2.5, fine-tuned LLama — 'ConflLlama') to human annotation.
- Compare the human coded dataset to what we get from the LLMs.

Generative LLM Prompt Training

To permit a comparison of model performance, we train generative LLMs via classification prompts. See Example

We use the GTD corpus for training and testing:

- Training texts: 1970 to 2016 (primarily data from 1998 to 2016)
- Test data texts and coded data: 2017 to 2020

Then use the predictions / outputs from Gemma, Llama, Qwen, etc. predict . . .

Performance Evaluation

We compare six language models for classifying terrorist events:

1. ConfliBERT: Domain-specific BERT model trained on conflict data
2. ConflLlama-Q4KM: Llama 3.1 (8B) fine-tuned on GTD data, 4-bit quantization
3. ConflLlama-Q8: Llama 3.1 (8B) fine-tuned on GTD data, 8-bit quantization
4. Gemma 2 (9B): Google's generative model with prompt training
5. Llama 3.1 (8B): Meta's generative model with prompt training
6. Qwen 2.5 (14B): Alibaba's generative model with prompt training

The quantized ConflLlama models (Q4KM, Q8) use reduced numerical precision to decrease memory usage while maintaining performance (Meher & Brandt, 2025).

Performance Evaluation

We evaluate models using metrics crucial for political event classification:

1. ROC curves - Assessing detection of conflict events
2. Accuracy - Overall event classification correctness
3. Precision - Avoiding false positives in conflict identification
4. Recall - Capturing all relevant political violence events
5. F1-Score - Balanced performance for skewed conflict data

These metrics are vital for building reliable political violence datasets that inform policy decisions. See simple case results BBC and re3d results

Preface Key Findings

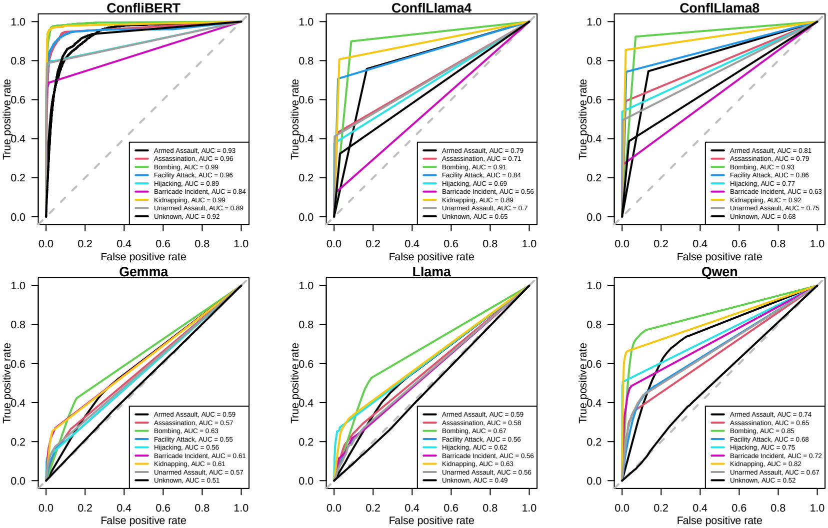
Performance Superiority: compared to Gen AI/LLMs,

- ConfliBERT achieves highest performance across tasks
- 150-200x faster on binary classification
- 300-400x faster on NER tasks
- Better precision-recall balance
- Domain-specific training outperforms larger models

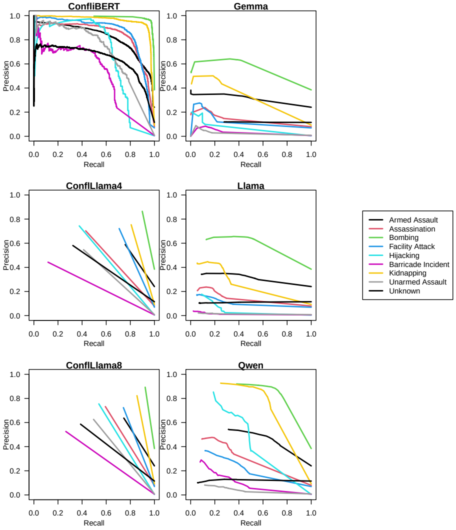
Counts of Predictions, 2017–2020

	GTD	ConfliBERT	CFLlama4	CFLlama8	Gemma	Llama	Qwen
Armed Assault	9079	10254	11686	10635	10467	10665	12072
Assassination	2990	2569	1830	2421	2565	2742	1953
Bombing	14508	14666	15089	15003	8017	9809	11011
Facility Attack	2624	3065	2574	2694	1313	3990	3507
Hijacking	154	125	78	110	91	660	110
Barricade Incident	230	118	61	116	560	1330	595
Kidnapping	3495	3214	3745	3633	1443	2072	2146
Unarmed Assault	301	255	227	238	694	2947	1798
Unknown	4328	3436	2410	2858	11902	4441	3807

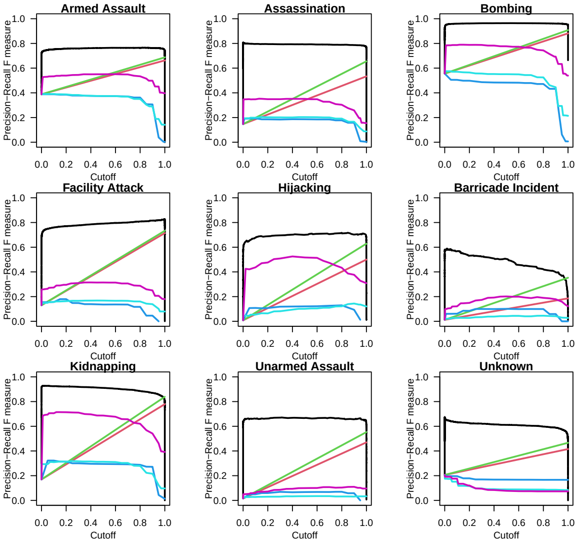
GTD Event ROCs and AUCs



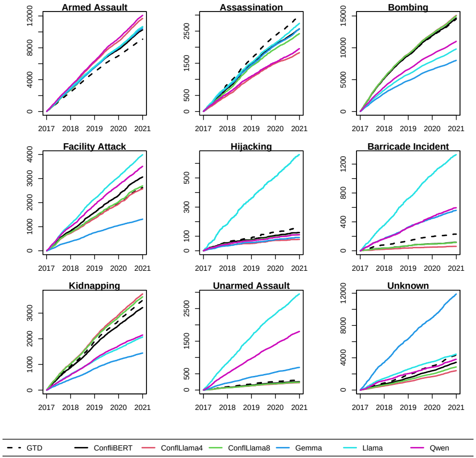
GTD Precision-Recall curves



GTD F SCORES


Armed Assault
Assassination
Bombing
Facility Attack
Hijacking
Barricade Incident
Kidnapping
Unarmed Assault
Unknown
Precision−Recall F measure
Cutoff
0.0
0.2
0.4
0.6
0.8
1.0

GTD Cumulative number of predicted events, 2017–2021 by type and model



Is it performative and fast?

Model	Accuracy	Precision	Recall	F1	Total Time	Time / Document	Relative Speed
ConfliBERT	0.90	0.83	0.77	0.79	27.6s	0.0016s	759.49x
ConflLlama-Q4KM*	0.72	0.72	0.72	0.71	49.9m	0.1746s	7.15x
ConflLlama-Q8*	0.76	0.76	0.76	0.75	52.3m	0.1831s	6.82x
Gemma 2	0.60	0.27	0.19	0.21	3.1h	0.6605s	1.89x
Llama 3.1	0.52	0.13	0.12	0.11	3.3h	0.7191s	1.74x
Qwen 2.5	0.74	0.50	0.44	0.45	5.8h	1.2490s	1.00x

Multi-Label Performance

Details are in the paper . . .

ConfliBERT Advantages:

- 79.4% subset accuracy - correctly identifies complex attacks
- 0.035 Hamming loss - lowest error rate
- Near-perfect label cardinality (0.907 vs 0.963) - captures event complexity
- Example: Syrian Civil War events combining armed assaults, bombings, and infrastructure attacks

Political Science Impact:

- Better conflict pattern recognition
- More accurate event complexity measurement
- Improved understanding of tactical combinations

Back of the envelope considerations

- You can begin this with a old codebook or small set of annotations (see Hu et al. 2024)
- This can be run on a someone powerful / recent desktop or laptop if you use the off the shelf model.
- We have trained the GTD example using conventional (non-HPC) hardware.
- This goes from a months and years $\rightarrow$ days and hours problems

Conclusions

- It is fast
- It is accurate, precise, etc.
- It is extensible and ready to be used

But remember:

- BERT is from 2017: it is in elementary school
- ChatGPT-alike is from 2022: it is a toddler

WHERE TO LEARN MORE ABOUT CONFLIBERT

Paper : `https://arxiv.org/abs/2412.15060`
Github : `https://github.com/eventdata/ConfliBERT`
Hugging Face: `https://huggingface.co/eventdata-utd`
Non-English versions:

- ConfliBERT in Spanish (es)
- ConfliBERT in Arabic (ar)
- Machine translation comparison : Osorio et al. (2024) ”Keep it Local: Comparing Domain-Specific LLMs in Native and Machine Translated Text using Parallel Corpora on Political Conflict”

See `https://eventdata.utdallas.edu/`

What do you need to bring to do this

1. Texts in digital form
2. Some labels or annotations of what you want coded or classified.
3. Can define a training-dev-test split across your texts? Can use and validate errors then, per Brandt and Sianan (2025, Frontiers in Political Science).

Other things we are working on or have thoughts about

- Other Gen AI?: No ChatGPT
- ConfliBERT in Spanish and Arabic Multilingual extensions
- Question-Answering approaches Some initial results Example 1 Example 2
- Machine translation $\rightarrow$ ConfliBERT? Parallel UN Corpora Paper / Keep it Local

Why Not Just Use ChatGPT?

Specialized vs. General Purpose:

- Speed: ConfliBERT 750x faster
- Cost: Local deployment vs API calls
- Control: Full access to model parameters
- Iteration: Rapid testing and refinement

Feature	ConfliBERT	General LLMs
Processing Time	Seconds	Hours
Deployment	Local	Cloud-based
Cost Model	One-time	Per-token
Customization	Full	Limited

Back

What is next?

- Multi-lingual comparisons: do you translate and then code, or build coders for each language?
- Dynamic network models to detect new actors and actions.
- Active learning for encoding information about these applications.
- Building new or extended datasets for say GTD (Duggan and Lafree), military exercises (D'Orazio → McManus and Nieman (2019), SNARP, MIDS, etc.
- MTL / LPC, Per Li et al. (2023, 2024).
- Life-cycle model development for updating and revising the models and training?
- ConflLlama: Llama 3.1 (8B) event coder!
- Other data sources....
- Question-Answer (QA) Applications:

Generative LLM Prompt

Example

Prompt: ”Classify each of the following events into up to three of these categories, providing probabilities for each: Assassination, Armed Assault, Bombing/Explosion, Hijacking, Hostage Taking (Barricade Incident), Hostage Taking (Kidnapping), Facility/Infrastructure Attack, Unarmed Assault, Unknown

For each event, return only a JSON object with category names as keys and probabilities as values.
Example format: {”Armed Assault”: 0.7, ”Bombing/Explosion”: 0.2, ”Unknown”: 0.1}

Back

Evaluation Data for BC & NER

Two Test Datasets:

- **BBC News Dataset**
  - 2,225 news articles
  - Binary labeled: conflict vs non-conflict
  - Diverse topics: business, politics, sports, tech
- **re3d Dataset**
  - Defense/security intelligence focus
  - Syria/Iraq conflict coverage
  - Expert-annotated entities (orgs, persons, locations)

Back

Model Performance Analysis

Model	Binary Class.		NER		Speed	
	Prec.	F1	Prec.	F1	Time(BC)	Time(NER)
ConfliBERT	**0.91**	**0.87**	**0.65**	**0.60**	**3.5s**	**1.4s**
Gemma 2	0.70	0.76	0.51	0.40	730.1s	866.2s
Llama 3.1	0.78	0.77	0.51	0.38	575.2s	489.4s

Back

ConfliBERT: Technical Development Pipeline

Core Technologies

- PyTorch + HuggingFace Transformers
- BERT base architecture (110M parameters)
- 8 NVIDIA GPUs for distributed training
- SimpleTransformers for fine-tuning

Training Data (33.7 GB)

- News articles (BBC, Reuters)
- Academic papers
- Policy documents
- Event databases (GTD, ACLED)
- Social media content

Output Variants

- ConfliBERT-scr-uncased (from scratch)
- ConfliBERT-scr-cased
- ConfliBERT-cont-uncased (continued)
- ConfliBERT-cont-cased

Binary Classification Example

Example

Input: "Two Lashkar e Jhangvi LeJ militants Asim alias Kapri and Ishaq alias Bobby confessed to killing four Rangers in Ittehad Town of Karachi, the provincial capital of Sindh."

Output: Gun Violence Related (1)

Input: "More than a week after a woman Communist Party of India-Maoist (CPI-Maoist) cadre was killed in an encounter in the forests of Lanjigarh block in Kalahandi District, the Maoists identified her as Sangita and called a bandh (general shutdown) in two Districts in protest against the killing."

Output: Gun Violence Related (1)

Back

NER Example

Example

Input: "A senior Muttahida Qaumi Movement (MQM) [ORG] worker identified as Sohail Rasheed [PERSON], 30, was shot dead near his home in Naeemabad [LOC] in Korangi Town [LOC] of Karachi [LOC], the provincial capital of Sindh [LOC], on June 19 [DATE]."

Output:

Perpetrator Organization: Muttahida Qaumi Movement (MQM)
Victim: Sohail Rasheed
Physical Target: Not specified
Location: Naeemabad, Korangi Town, Karachi, Sindh
Date: June 19

Back

Training Prompt Example

Example

Prompt: ”Classify each of the following events into up to three of these categories, providing probabilities for each: Assassination, Armed Assault, Bombing/Explosion, Hijacking, Hostage Taking (Barricade Incident), Hostage Taking (Kidnapping), Facility/Infrastructure Attack, Unarmed Assault, Unknown For each event, return only a JSON object with category names as keys and probabilities as values. Example format: {”Armed Assault”: 0.7, ”Bombing/Explosion”: 0.2, ”Unknown”: 0.1}
Events:”

(Multilingual) ConfliBERT Models

ConfliBERT ($>$ 33 GB text)

- Expert: United Nations, US State Department, NGOs
- English news: AP, PBS, NYT, Xinhua, AllAfrica
- Wikipedia: Topics for politics, government, war

ConfliBERT-Spanish ($>$ 30 GB text, 8.3 million documents)

- 123 news websites from 18 Spanish-speaking countries
- United Nations, European Union, 97 NGOs in 8 countries

ConfliBERT-Arabic ($>$ 30 GB text, 8.6 million documents)

- Primarily news from Arabic-speaking countries (e.g., Al Liwaa in Lebanon), including government news agencies
- Also Western sources like BBC Arabic, CNN Arabic

Back

(Multilingual) ConfliBERT Models

We have fine-tuned ConfliBERT models for:

- Binary classification
- Named entity recognition
- Multi-label classification
- Question-answering (prior to this, English only)

The goal of this research is to develop and test Question-Answering for the Spanish and Arabic models.

Question-Answering

Types of Question-Answering:

- Extractive: identifies the answer in a context without generating text. BERT is good at understanding content.
- Generative: uses the model to produce an answer. BERT is not as good for tasks involving text generation.

We focus on extractive QA because that is the process used to produce the data we use to study armed conflict.

- Armed Conflict Location and Event Data (ACLED)
- Militarized Interstate Dispute (MID)
- UCDP Georeferenced Event Data (GED)

Back

NYT, Patrick Kingsley and Euan Ward, “Live Updates: Wireless Devices Explore Across Lebanon After Israel Warns Hezbollah” 9/17/24

ConfliBERT

Select a task and provide the necessary inputs:

Select Task

Question Answering

Context

Large numbers of “wireless devices” simultaneously exploded across Lebanon in an apparently coordinated attack that caused hundreds of injuries, Lebanese health officials said on Tuesday, a day after Israeli leaders warned that they were considering stepping up their military campaign against Hezbollah.

Question

What is the conflict event?

Large numbers of “ wireless devices ” simultaneously exploded across Lebanon in an apparently coordinated attack

Submit

UTD Event Data | University of Texas at Dallas

NYT, Patrick Kingsley and Euan Ward, “Live Updates: Wireless Devices Explore Across Lebanon After Israel Warns Hezbollah” 9/17/24

ConfliBERT

Select a task and provide the necessary inputs:

Select Task

Question Answering

Context

Large numbers of “wireless devices” simultaneously exploded across Lebanon in an apparently coordinated attack that caused hundreds of injuries, Lebanese health officials said on Tuesday, a day after Israeli leaders warned that they were considering stepping up their military campaign against Hezbollah.

Question

Who is the target of the attack?

Hezbollah

Submit

UTD Event Data | University of Texas at Dallas

Back

Research Design

For training, QA datasets require: a Question, a Context (e.g., text of a story), and an Answer (span of text from story).

QA Datasets in Spanish

- NewsQA, translated from English to Spanish using the Translate Align Retrieve method, from CNN articles
- Spanish Question Answering Corpus (SQAC), texts in Spanish from Wikipedia, Wikinews, Newswire, AnCora

QA Datasets in Arabic

- XQUAD from Google Deepmind
- MLQA from Facebook Research
- ARCD from Arabic Wikipedia

RESEARCH DESIGN

Fine-tuned 8 models in Spanish:

- 4 ConfliBERT-Spanish (domain-specific corpora), initialized with BETO or BERT-multilingual vocabulary
- 2 BETO and 2 BERT-multilingual (generic corpora)

And 4 in Arabic:

- 2 ConfliBERT-Arabic (domain-specific corpora), initialized with AraBERT or BERT-multilingual
- 1 AraBERT and 1 BERT-multilingual (generic corpora)

All models fine-tuned with the same hyperparameters:

- 5 epochs, 5 different seeds, batch size 8, learning rate 5e-5

Results

TABLE: Results for Spanish

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

Results

Table: Results for Arabic

Model Name		(a) Extractive QA F1 Score	(a) Extractive QA Exact Match	(b) MLQA F1 Score	(b) MLQA Exact Match	(c) XQUAD F1 Score	(c) XQUAD Exact Match	(d) ARCD F1 Score	(d) ARCD Exact Match
ConfliBERT Arabic-v2	AraBERT	**61.90**	**40.11**	**64.86**	**44.24**	**63.33**	**47.19**	**57.43**	**28.92**
	Uncased	60.76	37.79	64.11	43.47	62.21	46.10	55.95	23.79
BERT	AraBERT	60.18	38.64	63.41	42.95	62.29	46.20	54.84	26.78
	Uncased	58.35	35.50	62.16	41.00	60.55	44.54	52.33	20.94

Example

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

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

"When did Hosni Mubarak take over the reins of power in Egypt?"

- ConfliBERT-Arabic: October, 1981
- BERT: 1950

Example

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

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

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

- ConfliBERT-Arabic: to the Supreme Council of the Armed Forces
- BERT: February 11, 2011

Additional Results

Experiments with ChatGPT:
We asked ChatGPT to "Answer questions based on the following text:" and then provided the context.
ChatGPT answered correctly: Mubarak came to power in Oct, 1981. But it added he did so after Anwar Sadat *resigned*.

- Anwar Sadat didn't resign, he was assassinated

We ran many tests, and in general ChatGPT had trouble with extractive QA. It couldn't help itself from generating stuff.

Original vs. Machine Translated (MT) Corpora

Conflict-related text typically is not gathered only in English but is collected in the native languages where the conflict occurs.

Do ConfliBERT variants perform better on native language or machine translated (MT) text?

MTs and Model Performance

We assess whether ConfliBERT's variants yield different results when processing MT text compared to native text.

- **Data Source:** 11,493 sentences from the UN Parallel Corpus (Ziemski et al. (2016) "The United Nations Parallel Corpus v1.0.") in English, Spanish, and Arabic as data source for comparison.
- **Annotations:** All sentences were coded for Binary (Relevant/ Non-Relevant) and QuadClass (Verbal/Material-Conflict/Cooperation) classification tasks by expert human coders.
- **Distribution:** The data consisted of 53.2% not relevant, 13.7% Material Conflict, 13.2% Material Cooperation, 8.3% Verbal Conflict, and 11.6% Verbal Cooperation sentences.

Research Design

- We translated native text using four commonly used MT tools (Google API, DeepL, Deep Learning, OPUS).
- We assessed MT quality using four quality metrics with differing flexibilty (BLEU, SacreBLEU, METEOR, BERTScore).
- We selected the best performing MT tool (DeepL) and assessed model performance in binary and multi-class classification tasks of different domain-specific and generic LLMs processing MT English text.

Performance Metrics MT text into English

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.

Performance Metrics MT text into English

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 *$p<0.1$, **$p<0.05$, ***$p<0.01$.

Native Language Performance across Languages

Next, we assess model performance across languages for both binary and multi-class classification.

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 * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Native Language Performance across Languages

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 * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

Comparison model performance MT vs. Native

Counterintuitively, we find that MT text yields better model performance results than native text with the exception of multi-class classification results for MT text from Spanish.

DIFFERENTIAL PERFORMANCE

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

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

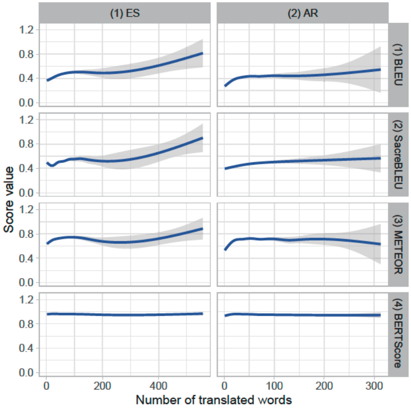
FINDING: MT-INDUCED VOCABULARY CHANGES

To better understand the performance improvement, we explore MT tool-induced changes to the native text. We find that:

- MT tools introduce heterogeneous changes in the data.
- DeepL both increases and decreases sentence-level word counts, depending on the source language.
- Word counts decrease for Spanish source text to English (-49,042 words/ -13.83%).
- Word counts increase from Arabic source text to English (+26,778 words / + 9.75

Word loss and MT quality

The MT-induced word loss affects MT quality metrics, leading to improvements or declines in quality score results. More succinct corpora are rewarded, more verbose corpora penalized.



MT tool-induced corpora changes and model perfromance

Building on the previous finding that MT tools lead to word count reductions and augmentations, we further assess the nature of these changes and their effect on ConfliBERT-variant model performance.

- **Across MT tools, which tools produce text that yields the best model performance compared to native text?**
- **What exactly is changed by MT tools and how do these changes affect model fit?**

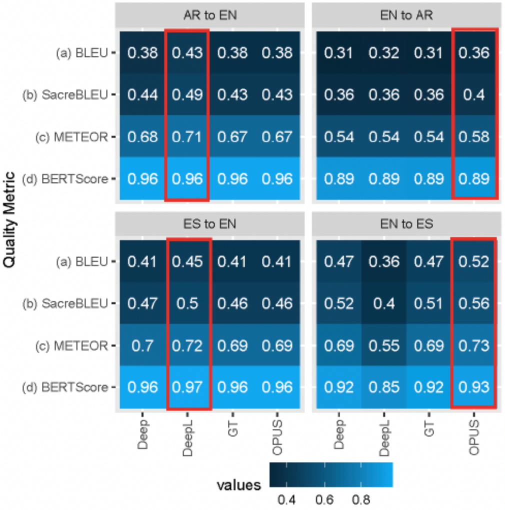
We continue to use the Ziemski et al. (2016) UN Parallel Corpus for our analyses.

MT quality in bidirectional translations

For this analysis, we conduct MTs from Arabic and Spanish into English, and vice versa, to assess MT tool performance in both directions.

- Bidirectional translations are conducted on Google Translate API, DeepL, Deep Learning, and OPUS.
- We continue to use BLEU, SacreBLEU, METEOR, and BERTScore as quality metrics.
- We find that DeepL yields the highest score for MT into English, while OPUS yields the highest score for translations into Spanish and Arabic.

MT TOOL QUALITY ASSESSMENTS

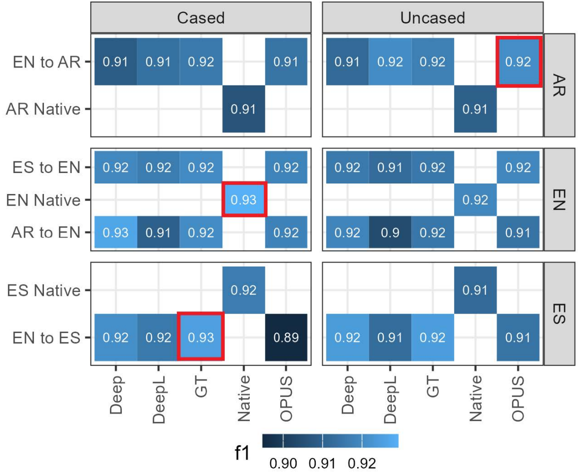


Model perfromance across MT tools

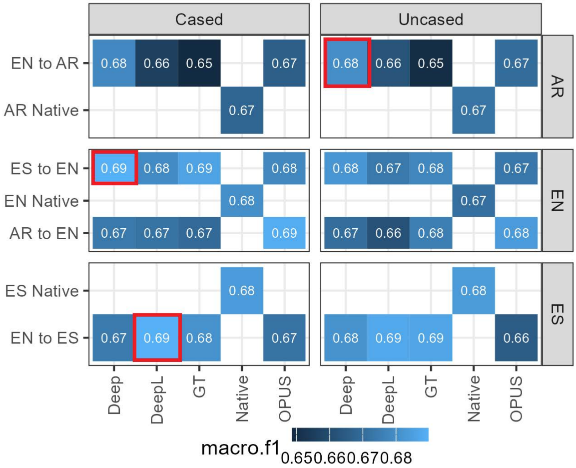
We then test the effect of MTs on LLM model performance.

- We conduct three classification tasks:
  - Relevant (Binary) classification
  - QuadClass (Multi-class) classification
  - BinQuad (Binary) classification of each QuadClass category
- We use all three ConfliBERT variants in cased and uncased variations resulting in six models.

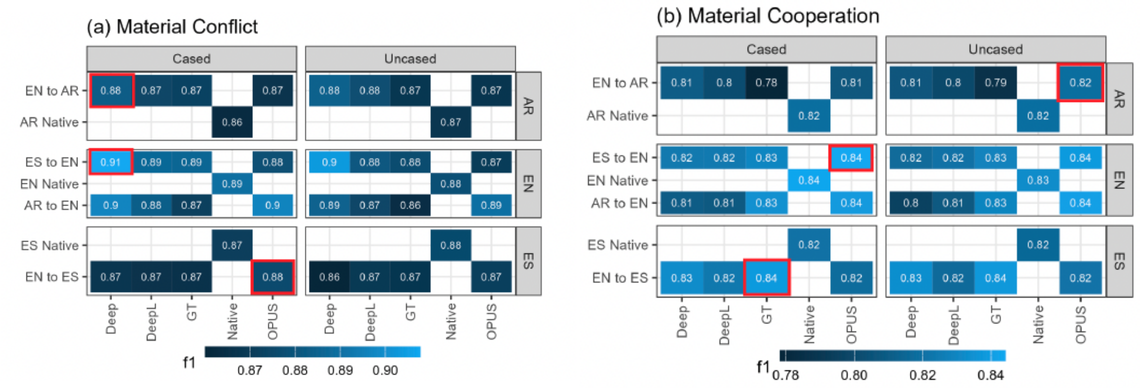
Binary Classification



QuadClass Classification



BinQuad Binary Classification - Material Conflict/Cooperation



BinQuad Binary Classification - Verbal Conflict/Cooperation

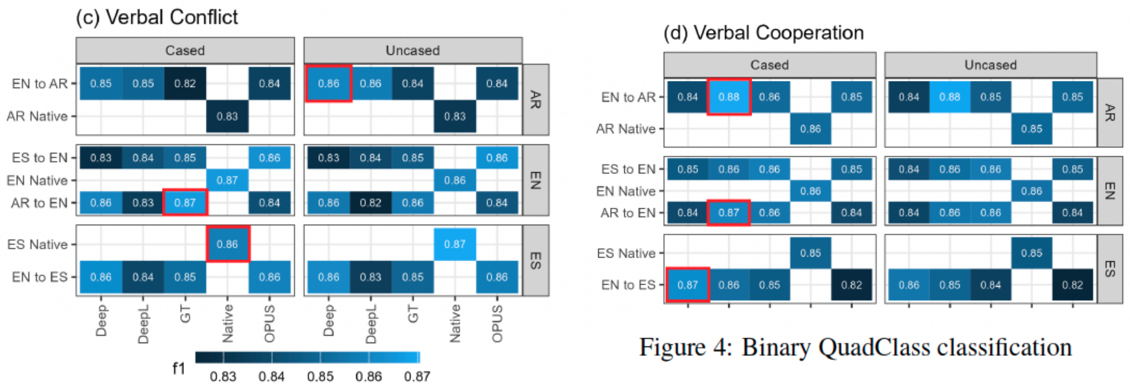
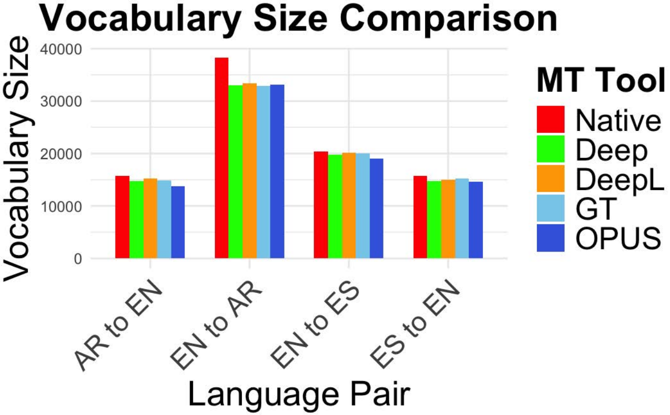


Figure 4: Binary QuadClass classification

MT-INDUCED VOCABULARY LOSS

To better understand MT-induced changes in the original corpora, we measure changes in vocabulary size for MT compared to the native corpora.



MT-induced loss in corpus rarity

We then measure text rarity per sentence, defining rarity as the proportion of tokens in a text that does not appear in the 5,000 most common tokens for a domain. We measure:

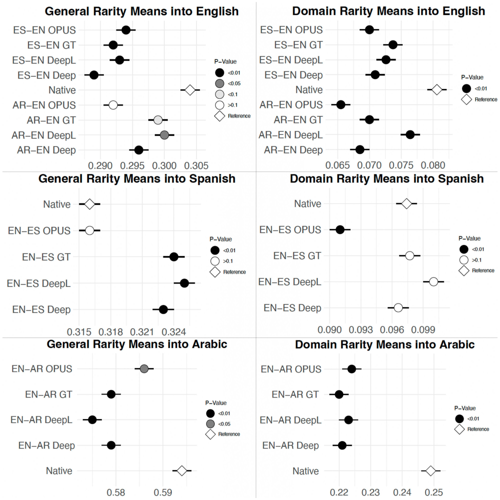
- General rarity: Relying on the 5,000 most common tokens for a language, regardless of the subject.
- Genre rarity: Relying on the 5,000 most common tokens in the sentences from the UN Parallel Corpus.

Rarity is used as a proxy for text complexity. A reduction of rarity them resembles a reduction of text complexity compared to the native corpus.

MT-induced loss in corpus rarity

- We find that English and Arabic MTs have lower general and domain rarity scores compared to native corpora.
- Spanish MTs have higher general rarity scores than the native corpus for Deep, DeepL, and GT. For domain rarity, the difference is not significant.

MT-induced loss in corpus rarity



Dependency Distance and sentence-level prediction confidence

- We further compare the Dependency Distance Mean between native an MT sentences as indicators of changes in sentence complexity.
- We also estimate the degree of confidence of ConfliBERT correctly classifying a sentence and explore determinants of model performance in the binary classification task.
- We evaluate the contribution of each variable on the probability of correct classification by comparing the contribution of each sentence-level characteristic to the regression Root Mean Standard Error (RMSE) using stepwise elimination. **We find that General and Domain Rarity scores lead to the largest model fit loss.**

Findings

- MT quality assessment scores provide limited insight about which MT tool performs best across classification tasks.
- MT tools induce a reduction in vocabulary complexity, leading to a loss of rare tokens that could be particularly relevant for domain experts.
- LLMs generally perform better with MT texts than with native corpora. There is no single MT tool that performs best across languages.
- **While machines talking to machines yields better results, this comes at a cost of losing richness and potentially relevant nuance in the MT.** Researchers considering using MT text over language specific LLMs need to consider this limitation.

Model Fit Loss by stepwise elimination