CPSS 2025

Proceedings of the 5th Workshop on Computational Linguistics for the Political and Social Sciences (CPSS)

Proceedings of the Workshop

September 10-11, 2025

The CPSS organizers gratefully acknowledge the support from the following sponsors.

GESIS Leibniz Institute
for the Social Sciences

GSCL
German Society
for Computational Linguistics
and Language Technology

©2025 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
`acl@aclweb.org`

ISBN None

Preface

It is our great pleasure to present the proceedings of the 5th Workshop on Computational Linguistics for the Political and Social Sciences (CPSS 2025), co-located with KONVENS 2025 in Hildesheim, Germany, held on September 10-11, 2025. CPSS is organized under the auspices of the GSCL Special Interest Group and serves as a vibrant interdisciplinary forum at the intersection of natural language processing, computational linguistics, and political and social science.

This year's workshop continues to bring together innovative research on text-as-data for the social sciences while spotlighting a special theme on validation and best practices for applying NLP techniques in political and social science research. Contributions addressed topics such as data quality in human and synthetic datasets, data leakage in large language models, dataset donation, and validation approaches beyond traditional train, development, and test paradigms.

We received 25 submissions in total. Following a rigorous double-blind peer review process, 10 papers were accepted for oral presentation, and 6 additional contributions were selected as non-archival posters. The archival papers will appear in this proceedings, and the non-archival abstracts showcase ongoing or early-stage work will be made available on the website and presented at the conference as posters.

Acknowledgements and sponsorship

Our thanks go to all authors for their excellent contributions, the reviewers for their thoughtful evaluations, and our invited speakers who inspired lively discussion. We also acknowledge the support from our sponsors GSCL, and GESIS - Leibniz Institute for the Social Sciences.

We extend our sincere gratitude to the KONVENS organizing committee for leadership and coordination, and to the program committee for their diligent commitment to rigorous and fair review.

The CPSS-2025 Co-Chairs:
Dennis Assenmacher (GESIS)
Christopher Klamm (University of Mannheim / CCCP, University of Cologne)
Gabriella Lapesa (GESIS / Heinrich-Heine University Düsseldorf)
Simone Ponzetto (University of Mannheim)
Ines Rehbein (University of Mannheim)
Indira Sen (University of Mannheim)

Organizing Committee

Chairs

Dennis Assenmacher, GESIS
Christopher Klamm, University of Mannheim / CCCP, University of Cologne
Gabriella Lapesa, GESIS / Heinrich-Heine University Düsseldorf
Simone Ponzetto, University of Mannheim
Ines Rehbein, University of Mannheim
Indira Sen, University of Mannheim

Program Committee

Committee Members

Aditi Dutta, University of Exeter
Ana Maria Lisboa dos Santos Cotovio, GESIS
Alexander Wuttke, LMU Munich
Andreas Küpfer, TU Darmstadt
Anna-Carolina Haensch, University of Munich
Annette Hautli-Janisz, University of Passau
Christian Arnold, Cardiff University
Chung-Hong Chan, GESIS
Clint Claessen, University of Basel
Daryna Dementieva, TUM Munich
Karolina Zaczynska, University of Potsdam
Lukas Erhard, University of Stuttgart
Max Pellert, University of Konstanz
Manfred Stede, University of Potsdam
Melanie Siegel, Hochschule Darmstadt
Sukayna Younger-Khan, University of Konstanz
Valerie Hase, LMU Munich
Verena Kunz, GESIS
Zlata Kikteva, University of Passau

Keynote Talk

From Psychometrics to Practice: Validating NLP for Political and Social Research

Qixiang Fang
Utrecht University
2025-09-11 09:00:00 – Room: **Bühler-Campus**

Abstract: The growing use of NLP in political and social science offers exciting opportunities, but also raises concerns about validity, reproducibility, and bias. In this talk, I discuss how psychometric principles can strengthen the evaluation of text-based measures, from construct validity in embeddings to benchmarking large language models with the PATCH framework. I also highlight challenges in reproducibility, showing how common flaws in human evaluation undermine trust in findings, and I reflect on the ethical risks of emerging applications such as personality inference. Building on these insights, I propose best practices to align NLP with social science standards. By integrating psychometric rigor with computational methods, we can make NLP a more reliable tool for understanding political and social phenomena.

Bio: Qixiang Fang is a postdoctoral researcher at Utrecht University and a senior member of the ODISSEI SoDa Team (`https://odissei-soda.nl/`). He advises social science and humanities researchers in the Netherlands on integrating NLP and other computational methods into their work. He organizes a recurring workshop on using large language models for data collection and on addressing measurement error in LLM-generated labels for downstream modeling (`https://sodascience.github.io/workshop_llm_data_collection/`). He obtained his PhD at Utrecht University, where his research explored how measurement theory can strengthen the application of NLP in the social sciences and humanities.

Keynote Talk

Measuring Political Bias in Large Language Models

Paul Röttger
Bocconi University

2025-09-11 11:00:00 – Room: **Bühler-Campus**

Abstract: Large language models (LLMs) are helping millions of users to learn and write about a diversity of issues. In doing so, LLMs may expose users to new ideas and perspectives, or reinforce existing knowledge and user opinions. This creates concerns about political bias in LLMs, and how such bias might influence LLM users and society. In my talk, I will discuss why measuring political bias in LLMs is difficult, and why we should be skeptical about most evidence so far. Then, I will present our approach to building a more meaningful evaluation dataset called IssueBench, to measure biases in how LLMs write about political issues. I will describe the steps we took to make IssueBench realistic and robust. Then, I will outline our results from testing state-of-the-art LLMs with IssueBench, including clear evidence for issue bias, striking similarities in biases across models, and strong alignment with Democrat over Republican voter positions on a subset of issues.

Bio: Paul is a postdoctoral researcher in the MilaNLP Lab at Bocconi University, working on evaluating and improving the alignment and safety of large language models, as well as measuring their societal impacts. For his recent work in this area, he won Outstanding Paper at ACL and Best Paper at NeurIPS D&B. Before coming to Milan, Paul completed his PhD at the University of Oxford, where he worked on LLMs for hate speech detection. During his PhD, Paul also co-founded a start-up building AI for content moderation, which was acquired by another large online safety company in 2023.

Table of Contents

Analyzing Polarization in Online Discourse on the 2023-2024 Israel–Hamas War
Daniel Miehling, Daniel Dakota and Sandra Kübler 1

How phatic is political communication in social media?
Aenne Cecilia Kristine Knierim, Aenne Cecilia Kristine Knierim, Ulrich Heid and Thomas Mandl 11

TRAVELWARN-Crawler: Constructing longitudinal datasets of government-issued travel warnings for political and social science research
Laura Braun and Christian Oswald 23

Causally fooled in the name of being honest? Evaluating causal extraction in LLMs for political text
Paulina Garcia-Corral, Hannah Dorothy Bechara, Krishnamoorthy Manohara and Slava Jankin30

Heroes, Villains, and Victims: Character Narratives in the WPS Agenda of the UNSC
Hannah Mathilde Steinbach, Imge Yüzüncüoglu, Raluca Rilla and Manfred Stede 48

Enhancing Policy Analysis with NLP: A Reproducible Approach to Incentive Classification
M.a. Waskow and John Philip McCrae 68

Narrative Structure Extraction in Disinformation and Trustworthy News: A Comparison of LLM, KG, and KG-Augmented Pipelines
Justina Mandravickaitė 80

Abstractive Event Analysis of Armed Conflicts: Introducing the UCDP-AEC Dataset
Étienne Simon, Helene Bøsei Olsen, Ramón Carreño, Rahul Mishra, Nikolay Arefyev, Mert Can Yilmaz, Lilja Øvrelid and Erik Velldal 98

Americans are dreamers – Generic statements and stereotyping in political tweets
Ines Rehbein 114

Program

Wednesday, September 10, 2025

09:00 - 09:15	*Welcome*
09:15 - 10:00	*Oral Session 1*
	Analyzing Polarization in Online Discourse on the 2023-2024 Israel–Hamas War Daniel Miehling, Daniel Dakota and Sandra Kübler
	Heroes, Villains, and Victims: Character Narratives in the WPS Agenda of the UNSC Hannah Mathilde Steinbach, Imge Yüzüncüoglu, Raluca Rilla and Manfred Stede
	Abstractive Event Analysis of Armed Conflicts: Introducing the UCDP-AEC Dataset Étienne Simon, Helene Bøsei Olsen, Ramón Carreño, Rahul Mishra, Nikolay Arefyev, Mert Can Yilmaz, Lilja Øvrelid and Erik Velldal
10:00 - 10:30	*Coffee*
10:30 - 11:30	*Oral Session 2*
	Americans are dreamers – Generic statements and stereotyping in political tweets Ines Rehbein
	Causally fooled in the name of being honest? Evaluating causal extraction in LLMs for political text Paulina Garcia-Corral, Hannah Dorothy Bechara, Krishnamoorthy Manohara and Slava Jankin
	How phatic is political communication in social media? Aenne Cecilia Kristine Knierim, Aenne Cecilia Kristine Knierim, Ulrich Heid and Thomas Mandl
	Enhancing Policy Analysis with NLP: A Reproducible Approach to Incentive Classification M.a. Waskow and John Philip McCrae
11:30 - 12:30	*Poster Session*
	Persona-driven Simulation of Voting Behavior in the European Parliament with Large Language Models Maximilian Kreutner, Marlene Lutz and Markus Strohmaier

What is Democracy? Exploiting Situation Entities to Uncover Democracy Frames in German Political Discourse
Julian Schlenker, Ines Rehbein and Simone Paolo Ponzetto

Are Epistemic Norms in Decline? A Comparative Analysis of Parliamentary Discourse
Segun Aroyehun, Fabio Carrella, Stephan Lewandowsky and David Garcia

How Stable Are Political Ideology Classifiers Over Time? An Empirical Evaluation of Temporal Robustness in NLP Models
Mohsin Khan

Leveraging Survey-Informed LLM Personas to Identify Cultural References : Enhancing Inclusivity in News Reporting
Reshmi Pillai, Wouter Van Atteveldt and Antske Fokkens

Thursday, September 11, 2025

09:00 - 10:00	*Keynote 1: Qixiang Fang (Utrecht University) From Psychometrics to Practice: Validating NLP for Political and Social Research*
10:00 - 10:30	*Coffee*
10:30 - 11:00	*Oral Session 3* *TRAVELWARN-Crawler: Constructing longitudinal datasets of government-issued travel warnings for political and social science research* Laura Braun and Christian Oswald *Narrative Structure Extraction in Disinformation and Trustworthy News: A Comparison of LLM, KG, and KG-Augmented Pipelines* Justina Mandravickaitė
11:00 - 12:00	*Keynote 2: Paul Röttger (Bocconi University) Measuring Political Bias in Large Language Models*
12:00 - 12:20	*Discussion: Validation and best practices for using NLP in political and social science research*
12:20 - 12:30	*Closing*

Analyzing Polarization in Online Discourse on the 2023-2024 Israel–Hamas War

Daniel Miehling, Daniel Dakota, Sandra Kübler
Indiana University
`{damieh, ddakota, skuebler}@iu.edu`

Abstract

We investigate large-scale sentiment analysis of YouTube Shorts in the context of the 2023-2024 Israel–Hamas war. Using a corpus of over 3 million user comments and replies from four state-funded or state-supported international media channels, we track sentiment toward key geopolitical and ideological entities over the course of one year. We investigate the correspondence between large-scale sentiment analysis and manual fine-grained analysis of trends as well as the correspondence between a longitudinal analysis and geopolitical events. Results show that overall sentiment trends depict user attitudes, but to interpret these patterns correctly, we need domain-specific knowledge and a combination of the automatic analysis with fine-grained manual analysis. We also show that the peaks of a longitudinal analysis correspond to (geo-)political events such as the Eurovision Song Contest.

1 Introduction

The October 7, 2023 attacks and the ensuing war between Israel and Hamas triggered a surge in global online discourse characterized by intense polarization and ideological fragmentation. While polarization in digital spaces has been widely studied, most analyses of political communication rely on coarse methods such as stance detection or user clustering, often applied to Twitter (Becatti et al., 2019; Stier et al., 2018). In contrast, we focus on using large-scale aspect-based sentiment analysis to investigate trends in user-generated content.

More specifically, we use aspect-based sentiment analysis (ABSA) to analyze over 3 million comments and replies on YouTube Shorts (a short-form video format similar to TikTok videos) by four state-funded or state-supported international broadcasters (Cull, 2008) (Al Jazeera, TRT World, BBC News, and Deutsche Welle)[1]. Since ABSA enables the detection of distinct sentiment toward multiple entities within the same utterance (Chauhan and Meena, 2019; Mai and Le, 2021), it can offer a nuanced view of ideological alignment. Consider the following example[2]:

(1) *Hamas being labeled as terrorist group is just a label. They are resistances against occupation and atrocities... Some countries like Israel and America have caused more dead a terror... but you can't label them Terrorists because they are countries.*

Here, *Hamas* is framed in a relativizing tone that challenges its designation as a terrorist group. In contrast, *Israel* and *America* are depicted as more violent actors, invoking negative sentiment. The final mention of *Terrorists* emphasizes a perceived double standard, adding a layer of ideological critique. Thus, ABSA allows these overlapping views to be disentangled.

Our study applies a current ABSA model (Yang et al., 2021, 2023) to investigate political communication in the digital mainstream. Drawing on insights from Kušen and Strembeck (2023), who show how emotional exposure during the early stages of the Ukraine war shaped user affect over time, we use large-scale sentiment analysis to map fine-grained sentiment toward competing and often opposing political actors. This approach enables us to track how emotional alignment and ideological positioning shift over time within digital discourse on YouTube. We trace sentiment trajectories over a 12-month period following the Hamas-led attack on Israel. We aim to examine the correspondence of sentiment analysis with fine-grained manual analysis as well as the correspondence to spikes in the longitudinal analysis to (geo-)political events such

---

[1] In our tables and figures, we use the abbreviations BBC and TRT to refer to BBC News and TRT World, respectively.

[2] The example represents user-generated content drawn from our corpus.

Proceedings of the 5th Workshop on Computational Linguistics for the Political and Social Sciences (CPSS), pages 1–10
September 10-11, 2025 ©2025 Association for Computational Linguistics

as the Eurovision Song Contest. More specifically, we ask the following research questions:

- **RQ1:** To what extent does large-scale sentiment analysis depict trends in user-generated content on Shorts videos?
- **RQ2:** How well do sentiment trends in a longitudinal analysis correspond to geopolitical events?

2 Related Work

With extensive prior research examining hate speech, misinformation, and political extremism in online spaces (Becker et al., 2023; Brown et al., 2024; Finkelstein et al., 2023; Rieger et al., 2020; Topor, 2024), sentiment analysis has emerged as a tool to model ideological alignment, polarization, and user engagement in socio-political discourse.

To capture the implicit and coded nature of political speech, Subramanian et al. (2023) and Young et al. (2024) emphasize the importance of context-sensitive sentiment models capable of interpreting nuance in polarized discourse. Shifts in media sentiment can precede conflict outbreaks (Jamison et al., 2023), with sentiment dynamics reflecting affective polarization (Lerman et al., 2024), and can capture emotional reactions in crisis moments (Kušen and Strembeck, 2023; Win Myint et al., 2024).

Aspect-based sentiment analysis has proven suitable for modeling longitudinal sentiment variation in politically charged online discussions. It captures more nuanced sentiment toward discrete entities or themes, enabling robust sentiment extraction at the aspect level across multiple targets within the same comment (Rietzler et al., 2020; Zhang et al., 2022), particularly in domains such as hate speech detection (Mughal et al., 2024; Zainuddin et al., 2016, 2018; Zhang et al., 2024) and political communication (Gold et al., 2018; Miok et al., 2023; Seno et al., 2024).

3 Methodology

3.1 Corpus and Collection

The empirical data was obtained via the official YouTube API[3]. We started the corpus collection two weeks after the events of October 7. To narrow the scope of this study, we focus on state-funded or state-supported international outlets: *TRT World* (Turkey), *BBC News* (United Kingdom), *Al Jazeera* (Qatar), and *Deutsche Welle* (Germany)[4]. These channels play a critical role in shaping public discourse and often reflect distinct geopolitical perspectives. Over a twelve-month data collection period, we gathered more than 3.4 million comments and replies from these four outlets, generated by over 780 000 unique users. Table 1 shows the total number of unique video IDs, user-generated comments and replies, and distinct users participating in discussions for each outlet.

Outlet	Videos	Comments	Users
DW	84	9 172	6 521
AJ	958	1 187 470	388 384
BBC	70	23 039	13 517
TRT	1 258	2 224 981	773 773
Total	2 370	3 444 662	787 157

Table 1: Corpus statistics across media outlets.

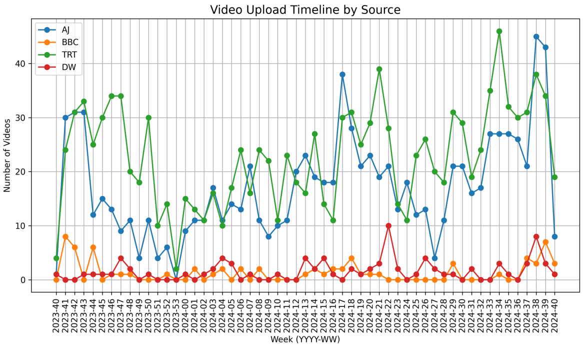


Figure 1: Weekly frequency of YouTube Shorts uploads per outlet over a 12-month period.

Figure 1 illustrates the weekly frequency of YouTube Shorts uploads per source. *TRT World* and *Al Jazeera* consistently released substantially more content, often exceeding 30 to 40 Shorts per week, compared to *BBC News* and *DW*, which uploaded fewer than 10 Shorts weekly. *BBC News* and *DW*, in turn, cover a broader thematic spectrum, with fewer videos directly addressing the Israel-Gaza conflict. While we initially considered including earlier data, a subsequent rescraping attempt on user-generated content related to BBC's Shorts revealed that nearly 50% of the original comments and replies had been removed, likely due to YouTube's content moderation policies.

---

[3] `https://developers.google.com/youtube/v3/docs`

[4] While *TRT World* has stated that it is not affiliated with the Turkish government, the other three channels are state-funded or state-supported broadcasters.

Outlet	Overall	Non-Eng.	Non-Eng. %
TRT	771 855	149 268	19.33
AJ	280 744	47 032	16.75
DW	2 733	386	14.12
BBC	10 298	850	8.25

Table 2: Remaining content after language filtering.

3.2 Preprocessing

We implemented a multi-step preprocessing pipeline to prepare the corpus for sentiment analysis. First, we filtered structurally invalid JSON files and removed entries with missing fields. We then normalized user mentions (e.g., `@username`) to support accurate tokenization. To ensure language consistency, we applied the `langdetect` library[5] and retained only English-language comments. For further quality control, we computed an English lexical coverage ratio using the `nltk.corpus.words` vocabulary[6] and discarded comments with less than 40% English terms (see Table 2).

To constrain the analysis to relevant discourse, we filtered the corpus using a targeted aspect lexicon that included geopolitical and ideological entities central to the conflict. This lexicon comprised the following terms: *Israel*, *Palestine*, *Jews*, *Palestinians*, *Zionists*, *Hamas*, *Hezbollah*, and *Muslims*[7].

3.3 Aspect-Based Annotation

Manual annotations were necessary to fine-tune an aspect-based sentiment analysis model to annotate the large scale corpus. The user-generated content in our corpus contains considerable noise, including unconventional spellings, typographical errors, informal language, and domain-specific allusions or “dog whistles”. To ensure accurate sentiment classification, we used a small subset of our corpus and annotated it for sentiment. We used `Label Studio`[8] (Tkachenko et al., 2025), an open-source data labeling platform, to annotate text segments with both aspect categories and sentiment labels. The annotation was conducted by an expert in the field. Sentences were pre-selected using dependency parsing (see Section 3.4) to identify those containing aspect terms from the aforementioned lexicon. Sentiment was annotated as *positive*, *neutral*, or *negative*, and applied to both explicit and clearly implied references. The annotation scheme accounted for informal syntax, metaphorical language, and ideological cues, which are characteristic of user-generated discourse.

Aspect	Negative	Neutral	Positive
Hamas	576	422	225
Hezbollah	5	14	12
Israel	296	278	376
Jews	192	244	109
Muslims	112	272	234
Palestine	175	186	101
Palestinians	272	261	260
Zionists	228	272	25
Total	1 856	1 949	1 342

Table 3: Distribution of annotated segments by aspect and sentiment label.

The final annotated corpus comprises 5 147 text segments[9], sampled from comments and replies across the four state-funded or state-supported media outlets. Selection criteria included lexical variety, comment diversity, and the removal of near-duplicate or semantically redundant content. We aimed to balance the sentiment distribution within each aspect category, although certain entities, particularly *Zionists*, were predominantly associated with negative sentiment, limiting class balance. Table 3 presents the distribution of annotated segments by aspect and sentiment label.

To assess reliability, a stratified random sample of 500 segments was independently annotated by a second expert. We computed inter-annotator agreement (IAA) on this subset, yielding scores of $\kappa = 0.86$ and $\alpha = 0.85$, indicating substantial agreement (see Table 4). While these scores reflect consistency, they do not guarantee conceptual accuracy in highly polarized or coded language domains (Paun et al., 2022), where ideological framings and implicit sentiment pose inherent challenges.

3.4 Dependency Parsing

We used the biaffine graph-based dependency parser (Dozat and Manning, 2017) implemented in

---

[5] `https://github.com/Mimino666/langdetect`

[6] `https://www.nltk.org/_modules/nltk/corpus/reader/wordlist.html`

[7] After processing, the aspect *Hezbollah* appeared only in a small number of examples and was therefore excluded from the ABSA evluation due to insufficient data.

[8] `https://labelstud.io`

[9] Each segment corresponds to a dependency-parsed sentence or clause containing at least one aspect mention. Due to informal punctuation, ellipses, or run-on constructions, some parsed segments span multiple clauses.

Sentiment	Cohen's κ	Krippendorff's α
Negative	0.850	0.850
Neutral	0.867	0.868
Positive	0.869	0.869
Overall	0.862	0.855

Table 4: Inter-annotator agreement scores across sentiment classes.

Class	Prec.	Rec.	F1
Negative	73.5	84.4	78.6
Neutral	78.4	71.9	75.0
Positive	85.0	77.9	81.3
Avg.	78.4	78.0	77.9

Table 5: Overall sentiment classification performance (on the test set).

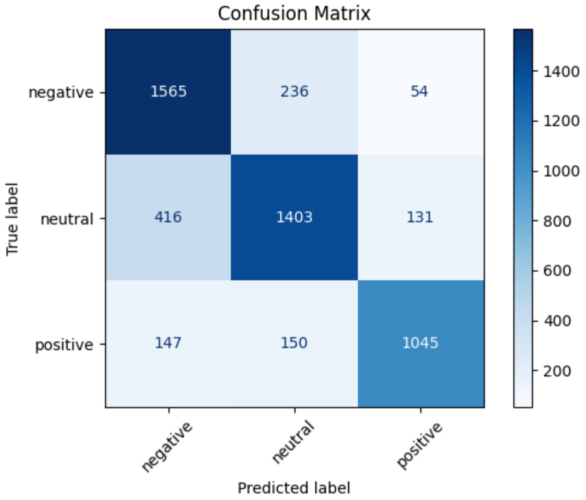


Figure 2: Confusion matrix for sentiment classification (on the test set).

the `SuPar` library[10] to extract syntactic structures, identify aspect terms, and locate their grammatical heads. In our workflow, the parser detected dependency relations within each comment, allowing us to align syntactic heads with predefined aspect categories for aspect term extraction (Zhang et al., 2022). Parsing was conducted using the pretrained `biaffine-dep-en` model, trained on English Universal Dependencies treebanks. We performed inference on a GPU with batch-wise processing over tokenized user-generated text.

3.5 Model Training and Evaluation

For aspect-based sentiment analysis, we used the end-to-end `DeBERTa-v3-large-absa-v1.1` model provided in the `pyABSA` library (Yang et al., 2021, 2023). This pretrained model was initially trained on English-language benchmark datasets, including Twitter and SemEval corpora, for aspect-based sentiment classification. We finetuned the model for 5 epochs on our task-specific training set using an NVIDIA A100 GPU. Training used a batch size of 2 and gradient accumulation over 8 steps, resulting in an effective batch size of 16. A cosine learning rate scheduler was used with an initial learning rate of 1×10^{-5}. As no separate development set was defined, we selected the best model checkpoint based on validation loss. Following the aspect-prompted classification framework, aspect terms in each sentence were replaced with a `$T$` marker. Input sequences were truncated to a maximum length of 512 tokens.

Table 5 shows the evaluation on the test set (precision, recall, and F1 across sentiment classes). The model achieves a macro-averaged F1 of 77.9, indicating strong performance in classifying explicit sentiment expressions related to our aspect terms.

The model most frequently confuses neutral comments and negative ones in both directions (see Figure 2). Since our goal is to assess sentiment directed toward specific ideological actors, we further evaluated per-aspect performance. As shown in Table 6, model performance is consistent across aspects, with particularly high accuracy for *Hezbollah*, *Muslims*, and *Palestine*, and solid results for more contentious aspect terms such as *Zionists* and *Hamas*. Our error analysis revealed three primary challenges: (1) difficulty in detecting implicit sentiment, particularly in sarcastic or coded language; (2) misclassifications in sentences containing ambiguous sentiment toward multiple aspects (e.g., "*Free Jews from Israel*"); and (3) challenges in processing code-switching comments and transliterations, where sentiment-bearing terms appeared in hybrid forms (e.g., Arabic-English transliterations).

4 Findings

To address our research questions, we adopt a two-step analytical approach that combines quantitative analysis with close qualitative reading. While ABSA enables large-scale sentiment classification, many findings cannot be interpreted in aggregation alone. In response to RQ1, we begin by analyz-

---

[10] `https://github.com/yzhangcs/parser`

Aspect	Acc.	Prec.	Rec.	F1
Hamas	75.1	75.0	75.1	74.6
Hezbollah	90.3	90.6	90.3	90.3
Israel	75.4	76.6	75.4	75.6
Jews	74.3	74.9	74.3	74.3
Muslims	85.4	85.8	85.4	85.5
Palestine	81.6	81.9	81.6	81.5
Palestinians	78.9	78.9	78.9	78.7
Zionists	79.0	79.6	79.0	78.7

Table 6: Per-aspect sentiment classification performance (on the test set).

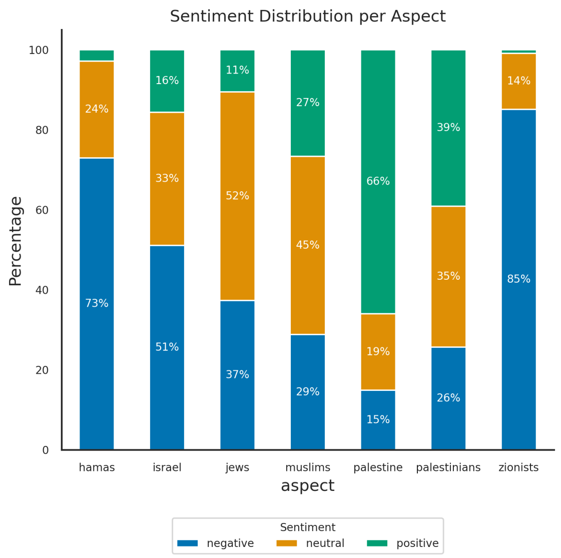


Figure 3: Stacked bar chart for sentiment distribution per aspect (in %).

ing the aggregated sentiment distribution across aspect terms to uncover general patterns of political communication in user-generated reactions to Shorts. We then turn to RQ2 by examining longitudinal sentiment dynamics and selected discourse patterns that reveal ideological alignments amid heightened volumes of user-generated content.

4.1 Aggregated Sentiment Distribution

Figure 3 presents the aggregated results of the sentiment analysis in form of the distribution of sentiment labels (negative, neutral, positive) across all classified aspect terms in the corpus, aggregated over all four media outlets. Negative sentiment dominates for the aspects *Israel*, *Zionists*, and, to a lesser extent, *Jews*. In contrast, *Palestinians* and *Muslims* receive more mixed or neutral sentiment, reflecting framings across the analyzed outlets. Interestingly, *Zionists* receive a higher proportion of negative sentiment (85%) than *Hamas* (73%), indicating that references to Zionism in mainstream YouTube discourse may elicit particularly adversarial responses[11].

While the aggregated results indicate a measurable share of positive sentiment toward *Palestine* and *Palestinians*, a closer examination of sentiment associated with *Jews* reveals that positive expressions do not necessarily reflect broadly favorable attitudes. Instead, such sentiment often appears in contrastive framings, for example, praising "good Jews" who oppose Israel or Zionism, while implicitly or explicitly condemning others, see examples (2) and (3). This discursive pattern reflects a form of conditional inclusion, where selective approval reinforces ideological boundaries. As mentioned earlier, such patterns of selective framing are not captured by sentiment proportions alone and require close reading to reveal how users shape ideological narratives.

(2) There are plenty of Jewish people fighting for Palestinian independence, we should not conflate Jewish people with the genocidal state of Israel.

(3) There are good orthodox Jews but Israel is run by a sadistic Zionist

A similar ambiguity emerges in references to *Palestinians*. In example (4), the model assigns positive sentiment, yet the phrase carries an implicit reference to a violent act (i.e., the "pager attack" linked to Hezbollah), thus revealing antagonistic intent beneath a lexically positive surface[12].

(4) Free Pagers for Palestine

Despite such conceptual ambiguities, these examples shed light on how evaluative language in user-generated content often carries multi-layered or contrastive meanings. In example (5), the phrase "Free Palestine" appears lexically positive, yet it is reframed as a critique of Hamas. Such contrastive uses illustrate how sentiment-laden expressions may simultaneously convey positive alignment with one political actor while denouncing

---

[11] While the proportion of negative sentiment is higher for *Zionists*, the total number of classified instances is much lower: 14 358 (*Zionists*) vs. 171 114 (*Hamas*).

[12] The phrase refers to events on 17–18 September 2024 in Lebanon, where explosives concealed in pager devices were deliberately detonated to target members of Hezbollah.

another. The presence of multiple aspect terms within short, syntactically simple comments suggests that such juxtapositions are common in political discourse and reveal how users employ sentiment cues to express ideological alignment in response to Shorts.

(5) Free Palestine from Hamas

While the majority of sentiment directed at *Hamas* is negative, we identified a smaller subset of positive sentiment. These instances often reflect implicit endorsement of Hamas as a resistance actor, as illustrated in example (6) and example (7).

(6) Victory to HAMAS-HEZBOLLAH FREE PALESTINE

(7) Kill every israel soldier who attack on falastin, hamas

Frequently, intensified sentiment is conveyed through symbolic and visual elements. Emoji sequences often accompany slogans like “Free Palestine” or denunciations of Zionism, that may serve to amplify affective tone or signal ideological stance. As shown in example (8), semiotic markers are combined to emphasize solidarity and resistance.

(8) FREE PALESTINE NOW! ✊🏽 PS✌ Down with 💤ionism and apartheid 🤮 🤬

Across the corpus, emojis are used extensively: 18.1% of all classified sentences (181 095 out of 1 000 750) contain at least one emoji. This high frequency highlights their role as affective intensifiers in discourse and underscores the need to consider multimodal cues as crucial elements of contemporary political communication within the digital mainstream. However, interpreting these patterns often requires domain-specific knowledge, as their meanings can diverge significantly from conventional language use.

4.2 Identifying Longitudinal Dynamics

To analyze how shifts in user discourse are reflected in sentiment patterns, we perform a peak analysis of weekly aspect-based sentiment trends. Figure 4 shows the overall frequency of sentiment-labeled content across the four state-funded or state-supported international broadcasters. It reveals key differences in user engagement: for instance, Shorts on *TRT World* consistently resulted in high volumes of user-generated content, while activity on *DW* remained sporadic, only increasing in response to specific events. We observe repeated spikes in negative sentiment that coincide with geopolitical events such as the Eurovision Song Contest or Israel’s Independence Day. However, instead of focusing solely on these general trends, we identify discursive tipping points: high aggregations of emotionally charged user responses that intensified user-generated reactions. We examine these sentiment surges more closely through aggregated data of weekly aspect-based sentiment trends, which allows us to trace how user sentiment fluctuated in response to unfolding real-world events.

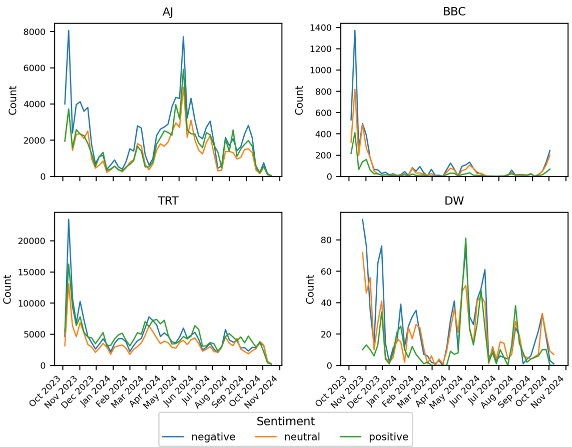


Figure 4: Sentiment trends over time across outlets.

Figure 5 shows sentiment trends per aspect. This analysis reveals three major peaks between March and May 2024, corresponding to heightened sentiment around *Palestine*, *Israel*, and *Zionists*. For each peak, we retrieved the most-commented Shorts videos and analyzed their content in relation to user responses. We compared sentiment curves across outlets and examined video titles and descriptions, together with comments and replies, to assess how user-generated activity reflects reporting on geopolitical events.

The first peak in early May coincides with the release of Shorts covering the Eurovision Song Contest, see example (9), and international campus protests, see example (10). High-engagement videos contributing to this spike included footage of Dutch riot police dismantling a Gaza protest camp and scenes of pro-Palestinian chants during Israel’s Eurovision rehearsal.

(9) Well done, to the audience at the Eurovision. Letting Israel know how the world feels about them !! FREE PALESTINE NOW !!!

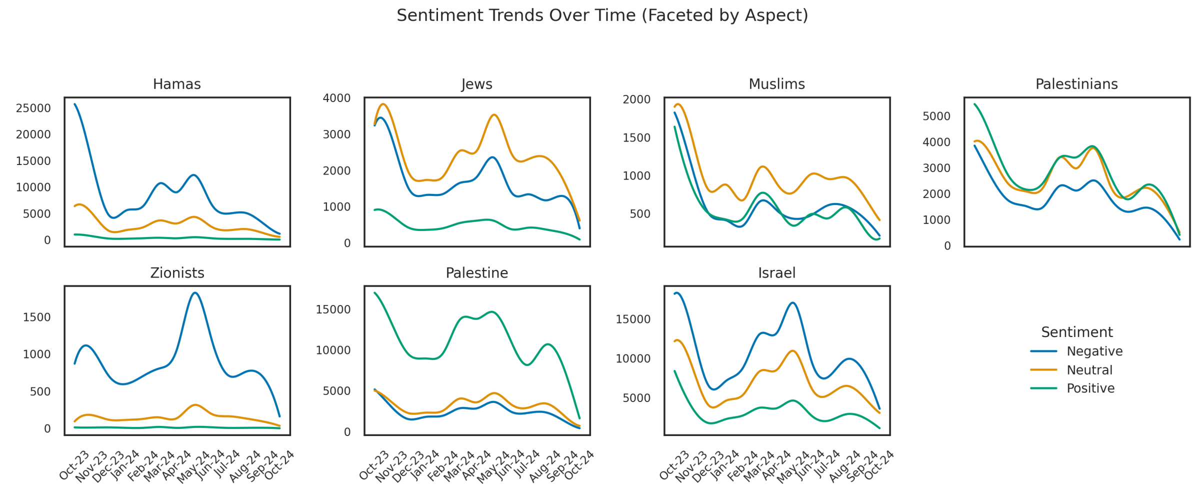


Figure 5: Sentiment trends over time across seven aspects.

(10) Many thanks the students who support majlum Palestine

These videos triggered waves of solidarity expressions such as “Free Palestine”, as well as accusations of genocide, see example (11), illustrating how users signaled political stance.

(11) Just need humanity to stand with Palestine. Stop genocide

The second peak coincided with Israel’s Independence Day on May 14 and was shaped by Shorts that cover allegations of the International Criminal Court (ICC) against Israel, as well as a Short depicting US-supplied weapons used in Gaza. During this period, negative sentiment toward *Israel* intensified. Users invoked narratives of indigeneity and dispossession, see examples (12) and (13), reflected by a high frequency of trigrams such as “Canaan Palestinians descendants” and “indigenous people Palestinians” to assert historical and moral claims.

(12) Wasn’t Israel made up by the whiteEuropeans, Delete the Hebrews field of land from the indigenous canaanites?

(13) The indigenous people are the Palestinians not the israeles who’ve been separated from Canaan and the old israel for a very long time

A third peak followed shortly thereafter, marked by intensified negative sentiment toward *Zionists*. This spike coincide with Shorts depicting police crackdowns on protest camps and a widely viewed video showing a scholar harassing a Muslim woman at a pro-Israel protest. Commenters responded with ideologically charged language, adopting anti-colonial framings and linking protest repression to broader narratives of systemic violence. Phrases such as “Zionists are dangerous”, see example (14), and “Zionist colonizers”, see example (15), highlight how the discourse shifted from solidarity-based expressions to overt ideological opposition.

(14) Zionists are dangerous for all people of the world.

(15) You mean Zionists hate them spreading the truth and standing up against the evil Zionist colonizers committing genocide.

Compared to earlier peaks, the third peak features significantly more pejorative and antagonistic language. Users adopt moral binaries and dehumanizing rhetoric to frame the conflict between Hamas and Israel, shifting from solidarity-based expressions to ideologically charged attacks targeting *Zionists*, marking a clear intensification of discursive polarization.

Taken together, the comparison across these peaks suggests a transition from expressions of solidarity to increasingly polarizing and antagonistic rhetoric. User-generated content toward *Zionists* and *Israel* exhibits more sentiment shifts than content related to *Palestine*. Users responded to geopolitical flashpoints by asserting legitimacy claims, voicing collective grievances, and expressing identity-based solidarities, dynamics consistent with prior research on event-driven polarization (Royesh and Grossman, 2021; Miehling, 2024; Alamsyah et al., 2024; Kušen and Strembeck, 2023).

5 Discussion

Our findings suggest that sentiment in user-generated content functions not merely as an emotional signal, but as a means of ideological expression and alignment. ABSA enables us to trace how users position themselves in relation to geopolitical actors through coded language and contrastive framings.

These patterns were especially salient in discourse surrounding *Zionists* and *Israel*, which were frequently associated with accusations of colonialism, genocide, and systemic violence. In contrast, while the aspect *Hamas* displayed only a small share of positive sentiment, its occasional framing in terms of resistance or heroism illustrates how affective discourse can operate through implication and conceptual ambiguity rather than overt endorsement. In particular, *Zionists* received the highest proportion of negative sentiment, even exceeding *Hamas*. While this trend underscores adversarial positioning toward Zionism as an ideology, the lack of positive examples also posed challenges for classification. To maintain empirical transparency, and to avoid biasing the ABSA system, we refrained from augmenting the training data, e.g., by generating synthetic data, to address class imbalance.

At the same time, our findings demonstrate the utility of large-scale sentiment analysis not only for identifying sentiment polarity, but also for unpacking rhetorical strategies users deploy to articulate political stances. In this context, affect operates discursively, conveying moral evaluation and ideological positioning. We observed persistent ambiguity in contrastive expressions involving Jews, where seemingly positive portrayals of Jewish fringe groups often legitimized the simultaneous condemnation of Israel or *Zionists*. Such constructions not only complicate the interpretation of straightforward sentiments but also reflect deeper ideological divisions.

Temporal sentiment patterns in our peak analysis further reveal that user-generated content extends beyond immediate reactions to breaking news, popcultural events, or protest repression. Affective intensification often mirrors geopolitical flashpoints but also transcends them. The accompanying discourse frequently moves beyond the specific Short-video context. Users employ emotionally charged, morally coded language to express solidarity, outrage, and condemnation. Thus, user-generated content should not be understood solely as real-time responses, but as discursive echoes, shaped by collective memory and entrenched resentment. Taken together, these findings suggest that sentiment analysis leads to meaningful results for interpreting political discourse online, not as a series of isolated opinions, but as structured, emotionally mediated ideological formations.

6 Conclusion and Future Work

We have investigated the interconnection of sentiment and polarization in political discourse. To answer RQ1, we have demonstrated that large-scale sentiment analysis can capture patterns of political communication embedded in user-generated discourse. While surface-level trends reveal general sentiment, a closer reading of aspect-specific results, such as contrastive framings (“good Jews” vs. “Zionists”), semiotic markers, and selective sentiment toward ideologically loaded terms, shows how users employ affective language to articulate political stance, moral judgment, and group alignment.

Addressing RQ2, we showed that we can identify evolving discourse patterns and ideological shifts in response to real-world geopolitical events. By analyzing sentiment peaks and comparing aspect-specific dynamics, we traced users expressing both solidarity and polarization, revealing affective intensification and discursive realignment over time.

For future work, we will extend our investigation and analyze more sentiment targets (aspects). We will also investigate the usability of an RST parser to gain deeper insights into how the discourse is structured beyond the sentence level.

Limitations

Parsing: While dependency parsing enhanced aspect-term extraction, it did not fully resolve ambiguities in sentiment attribution. Future improvements may benefit from incorporating contrastive learning techniques or domain-adapted embeddings to better capture context-sensitive sentiment in highly polarized political discourse.

Annotations: Some statements lack clear intent markers (e.g., sarcasm, irony, rhetorical questions). We infer meaning based on domain-specific expertise using established patterns of dog whistles and coded language, but some ambiguous cases remain unresolved.

Corpus Constraints: Several initial aspects (e.g., *Hezbollah*) were discarded because many comments failed the language threshold ($\geq$ 40% English), indicating widespread mixed-language use. Other aspects (e.g., *Zionists*) had very few positive examples, which led to class imbalance that impacted model performance. While augmenting the corpus could mitigate some of these limitations, generating synthetic samples—particularly in sensitive political domains—raises ethical concerns, despite the potential to improve accuracy.

The results on user-generated content from *Deutsche Welle (DW)* and *BBC News* should be interpreted with caution: due to limited short-video content published by these outlets, our corpus contains significantly fewer samples from these sources compared to *TRT World* and *Al Jazeera (AJ)*. This discrepancy does not imply that *DW* or *BBC News* did not cover the aftermath of October 7, but rather reflects platform-specific publishing practices.

Ethics Statement

Our study was conducted in accordance with institutional IRB approval and ethical research standards. We show user-generated content that includes sensitive and potentially harmful language, such expressions that may endorse or glorify violence. These examples are necessary to address our research questions and serve solely for illustrative purposes.

Acknowledgments

This work used Jetstream2 at Indiana University through allocation HUM200003 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

Andry Alamsyah, Akmal Wijaya Muharam, and Dian Puteri Ramadhani. 2024. Polarized narratives in digital spaces: A social network examination of the Gaza conflict. In *2024 International Conference on Data Science and Its Applications (ICoDSA)*, pages 527–532.

Carolina. Becatti, Guida Caldarelli, Renaud Lambiotte, and Fabio Saracco. 2019. Extracting significant signal of news consumption from social networks: The case of Twitter in Italian political elections. *Palgrave Commun*, 5(91).

Matthias J. Becker, Laura. Ascone, Matthew Bolton, Alexis Chapelan, Pia Haupeltshofer, Alexa Krugel, Karolina. Placzynta, Marcus Scheiber, and Victor Tschiskale. 2023. Celebrating terror: Antisemitism online after the Hamas attacks on Israel.

Olivia Brown, Laura G. E. Smith, Brittany I. Davidson, Daniel Racek, and Adam Joinson. 2024. Online signals of extremist mobilization. *Personality and Social Psychology Bulletin*.

Ganpat Singh Chauhan and Yogesh Kumar Meena. 2019. Youtube video ranking by aspect-based sentiment analysis on user feedback. In J. Wang, G. R. M. Reddy, V. K. Prasad, and V. S. Reddy, editors, *Soft Computing and Signal Processing*, volume 900, pages 63–71. Springer Singapore.

Nicholas J. Cull. 2008. Public diplomacy: Taxonomies and histories. *The Annals of the American Academy of Political and Social Science*, 616(1):31–54.

Timothy Dozat and Christopher Manning. 2017. Deep biaffine attention for neural dependency parsing. In *5h International Conference on Learning Representations (ICLR 2017)*, Toulon, France.

Joel Finkelstein, Jack Donohue, Alex Goldenberg, Lee Jussim, Alexander Reid Ross, Collin Vasko, Christian Ramos, Taylor Glover, Anisha Jagdeep, Anthony Boyle, and Jiya Modi. 2023. #Twittertakeover: How the Musk acquisition became a launchpad for Gen-Z Neo-Nazis, Ye, and widespread antisemitism. Technical report, Network Contagion Research Institute.

Darina Gold, Marie Bexte, and Torsten Zesch. 2018. Corpus of aspect-based sentiment in political debates. In *Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)*, pages 89–99, Vienna, Austria.

Ashley S Jamison, Jason Miklian, Kristian Hoelscher, Witold J Henisz, and Brian Ganson. 2023. Is media sentiment associated with future conflict events? *SSRN Electronic Journal*.

Eduard Kušen and Mark Strembeck. 2023. The effects of multiple exposure to highly emotional social media content during the early stages of the 2022 war in Ukraine. *SN Computer Science*, 4(1):663.

Kristina Lerman, Dan Feldman, Zihao He, and Ashwin Rao. 2024. Affective polarization and dynamics of information spread in online networks. *npj Complexity*, 1(8).

Long Mai and Bac Le. 2021. Joint sentence and aspect-level sentiment analysis of product comments. *Annals of Operations Research*, 300(2):493–513.

Daniel Miehling. 2024. *Online-Antisemitismus verstehen: Hassrede im Web 2.0. Eine vergleichende Studie zwischen der Mainstream-Plattform X und Fringe-Communities auf Telegram*, 1 edition. Nomos Verlag, Baden-Baden.

Kristian Miok, Encarnacion Hidalgo-Tenorio, Petya Osenova, Miguel-Angel Benitez-Castro, and Marko Robnik-Sikonja. 2023. Multi-aspect multilingual and cross-lingual parliamentary speech analysis. *arXiv preprint arXiv:2207.01054*.

Nimra Mughal, Ghulam Mujtaba, Sarang Shaikh, Aveenash Kumar, and Sher Muhammad Daudpota. 2024. Comparative analysis of deep natural networks and large language models for aspect-based sentiment analysis. *IEEE Access*, 12:60943–60959.

Silviu Paun, Ron Artstein, and Massimo Poesio. 2022. *Statistical Methods for Annotation Analysis*. Springer International Publishing, Cham.

Diana Rieger, Lena Frischlich, Stefanie Rack, and Gary Bente. 2020. Digitaler Wandel, Radikalisierungsprozesse und Extremismusprävention im Internet. In *Handbuch Extremismusprävention – Gesamtgesellschaftlich. Phänomenübergreifend*, pages 351–388. Bundeskriminalamt.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4933–4941, Marseille, France. European Language Resources Association.

Abuzar Royesh and Shelby Grossman. #COVID1948: How a hashtag on Israel and Palestine went viral [online]. 2021. Freeman Spogli Institute for International Studies, Stanford University.

Eloize Seno, Lucs Silva, Fábio Anno, Fabiano Rocha, and Helena Caseli. 2024. Aspect-based sentiment analysis in comments on political debates in Portuguese: Evaluating the potential of ChatGPT. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR)*, pages 312–320.

Sebastian Stier, Arnim Bleier, Haiko Lietz, and Markus Strohmaier. 2018. Election campaigning on social media: Politicians, audiences, and the mediation of political communication on Facebook and Twitter. *Political Communication*, 35(1):50–74.

Malliga Subramanian, Veerappampalayam Easwaramoorthy Sathiskumar, G. Deepalakshmi, Jaehyuk Cho, and G. Manikandan. 2023. A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. *Alexandria Engineering Journal*, 80:110–121.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Lev Topor. 2024. Mis/disinformation and national resilience: Are countries immune to fake news? In *Cyber Sovereignty: International Security, Mass Communication, and the Future of the Internet*, pages 111–131. Springer Nature Switzerland.

Phyu Yi Win Myint, Sin Lo Lo, and Yanchun Zhang. 2024. Unveiling the dynamics of crisis events: Sentiment and emotion analysis via multi-task learning with attention mechanism and subject-based intent prediction. *Information Processing & Management*, 61(4):103413.

Heng Yang, Biqing Zeng, Mayi Xu, and Tianxing Wang. 2021. Back to reality: Leveraging pattern-driven modeling to enable affordable sentiment dependency learning. *CoRR*, abs/2110.08604.

Heng Yang, Chen Zhang, and Ke Li. 2023. PyABSA: A modularized framework for reproducible aspect-based sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM ’23, page 5117–5122, New York, NY, USA. Association for Computing Machinery.

James C. Young, Rudy Arthur, and Hywel T. P. Williams. 2024. CIDER: Context-sensitive polarity measurement for short-form text. *PLOS ONE*, 19(4):e0299490.

Nurulhuda Zainuddin, Ali Selamat, and Roliana Ibrahim. 2016. Improving Twitter aspect-based sentiment analysis using hybrid approach. In *Intelligent Information and Database Systems*, pages 151–160. Springer.

Nurulhuda Zainuddin, Ali Selamat, and Roliana Ibrahim. 2018. Hybrid sentiment classification on Twitter aspect-based sentiment analysis. *Applied Intelligence*, 48(5):1218–1232.

Hao Zhang, Yu-N Cheah, Osamah Mohammed Alyasiri, and Jieuyu An. 2024. Exploring aspect-based sentiment quadruple extraction with implicit aspects, opinions, and ChatGPT: A comprehensive survey. *Artificial Intelligence Review*, 57(2):17.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges (version 2). *arXiv preprint arXiv:2203.01054*.

How phatic is political communication in social media?

Aenne C. K. Knierim and **Jannis Kuck** and **Ulrich Heid** and **Thomas Mandl**
Department of Information Science and Natural Language Processing
University of Hildesheim
Universitätsplatz 1, 31141 Hildesheim, Germany
`{knierim,kuck,heidul,mandl}@uni-hildesheim.de`

Abstract

This paper explores the role of phatic communication in political discourse on social media. We present a case study of German political communication on Instagram in the context of protests against antisemitism and the rise of right-wing extremism in Germany (October 2023 to April 2024). Focusing on non-argumentative posts, we annotate the data according to three categories: phatic, mobilizing, and informative. To assess the classification of phatic content, we compare human annotation with the performance of large language models (LLMs) in both zero-shot and few-shot classification settings. Inter-annotator agreement among human experts is moderate (Cohen's K = 0.62), highlighting the inherent ambiguity in identifying phatic communication. Among the models evaluated, a Mistral-based model achieves the best performance. Our findings suggest that while phatic elements are present in political discourse, the majority of posts serve mobilizing or informative functions. Our analysis illustrates, by way of example, that very complex annotations can only be performed by models with limited quality. Depending on the class and task, heterogeneous results are obtained

1 Introduction

The notion of phatic communion was introduced by Malinowski (1967) and describes language that is used exclusively to fulfill a social function and without conveying information: "Each utterance is an act serving the direct aim of binding hearer to speaker by a tie of some social sentiment or other". Phatic communication is omnipresent in social media, be it through status updates like "currently eating ice cream" or phatic photo sharing (Niemelä-Nyrhinen and Seppänen, 2020). Some even argue social media are phatic media, in which communication without content has taken precedence (Miller, 2008). This, according to Miller (2017), has led to social media as environments based on "limited forms of expressive solidarity as opposed to an engaged, content-driven, dialogic public sphere".

Social networks are full of informal, nonargumentative texts. Phatic media, or technologies, help sustain relationships through pervasive but non-informational contact (Miller, 2008). The constant stream of communication generated by smartphones and the availability of social media encourages a type of communication that suggests a general sociability and affability without exchanging real information (Miller, 2017).

As many texts are informal and media have been labeled to be phatic, scholars debate the relevance of political communication in social media. Without doubt, the so-called micro-activism, slacktivism, or clicktivism has been studied extensively: More than 300 studies have used survey data to test the relationship between the use of digital media and offline engagement in civic and political life, such as voting, volunteering, and protesting (Boulianne, 2020). This impact was deduced from a statistical synthesis by Boulianne (2020). Critical voices have challenged "digital politics enthusiasts" and ask if political communication has become increasingly phatic and less dialogic (Miller, 2017).

Undoubtedly, phatic communication is prevalent in social media. But is it also prevalent in political communication in social media, as in new social movements or activist dicsourse? In this paper, we want to explore what the real amount of phatic communication is in political communication. We present a case study analyzing a German Instagram corpus structured by the hashtags #WirSindMehr and #NieWiederIstJetzt. In previous studies (Knierim and Heid, 2025; Knierim et al., 2025), argumentation in this discourse has been investigated. In this paper, we investigate a corpus of non-argumentative sentences, following the premise that they are more likely to be phatic. The con-

Proceedings of the 5th Workshop on Computational Linguistics for the Political and Social Sciences (CPSS), pages 11–22
September 10-11, 2025 ©2025 Association for Computational Linguistics

tributions of this paper are three-fold: First, we put forward annotation guidelines and present a dataset of potentially phatic texts in German. We compare the performance of few-shot and zero-shot-settings of two large language models on detecting phatic communication. Third, we demonstrate the distribution of phatic content in our dataset to quantify the amount of phatic communication in an example of political online discourse.

In this paper, we first introduce concepts by researchers that try to grasp phatic language in social media. We then describe the corpus compilation and annotation process. Then, the prompt design and classification task are described. Finally, we present the results: we first discuss the distribution of phatic language in our corpus. Then, we discuss the performance of large language models for this task.

2 Phatic communication in social media

Microposts and phatic posts are derivatives of phatic communication (Radovanovic and Ragnedda, 2012). Although they might not have information value (Malinowski, 1967), Radovanovic and Ragnedda (2012) argue that phatic posts do have semantic and social value for the interlocutors. Following Malinowski (1967), Radovanovic and Ragnedda (2012) identify functions of phatic communication: Firstly, phatic communication serves the purpose of “social upkeep”. Secondly, phatic posts can signal the availabilty as a potential communication partner. Thirdly, phatic posts have a validation and recognition function to indicate recognition of one’s interlocutor as a potential communicative partner.

Radovanovic and Ragnedda (2012) argue that modern-day micro posts have their origin in the human need for social upkeep. They related phatic posts to the phatic function introduced by Jakobson (1968). Jacobson introduced the five general functions of language, among them the phatic function (try to maintain in contact with the receiver). Radovanovic and Ragnedda (2012) describe phatic posts as “the phatic display of connected presence”, where users engage through brief, informal messages. Following the authors, this signal also serves to validate the interlocutor as a potential communication partner.

The constant stream of communication created by smartphones and the availability of social media encourages a type of communication that suggests a general sociability and affability without exchanging real information (Miller, 2017). Miller (2017) challenges the notion that the ability of technology to connect people necessarily leads to an activism that involves dissent or resistance: Specifically, “political talk” would express convivial solidarity with others and have a function of social inclusion and recognition. Political communication in social media promotes phatic or idle talk, instead of working towards transformational political action (Miller, 2008). In this paper, we address this research question from a quantitative viewpoint by measuring the proportion of phatic communication in non-argumentative texts, following the premise that these are more likely to be phatic than argumentative posts.

Sarjanoja et al. (2013) conduct interviews to analyze the content of status updates on Facebook. They consider a sample of N=484 unique status updates, with participants rating how interesting they find them. The status updates are manually classified into topical categories, such as “small talk”, “current events talk” and “complaining”. Their findings indicate that Facebook status updates have strong elements of phatic communication and are used to increase one’s social capital (Sarjanoja et al., 2013). Interestingly, participants found most status updates to be “boring”. This demonstrates how phatic communication can be considered to be mundaine while still serving a social function. In a study about phatic photo sharing on Snapchat, Rettberg (2018) argues that the phatic connection between people is the key element, “often far more important than the content that is shared”. This relates to the non-archival character of some platform affordances, translating to increased immediacy and a sense of shared experience (Rettberg, 2018).

Duffy and Ling (2020) study the sociality of news sharing on social media, conducting a survey of N=2000 participants and using data from 88 focus groups. Participants’ motivations in sharing news include being part of a conversation, maintaining friendships, and contributing to a group. Not all news sharing is phatic according to the authors, as it can serve to warn, to inform or to advise. The authors argue that “sharing news online often conforms to the phatic, non- directional conversational gambit of Malinowski, and one focus-group participant said that people share the news because *’[when] there is no topic to start a conversation, we start a conversation on the news’* ”.

Phatic communication was initially studied in language but research has shown that it also occurs in photo-sharing (Niemelä-Nyrhinen and Seppänen, 2020) and even in the form of simple "likes" and reactions to others' posts (Radovanovic and Ragnedda, 2012). For example in snapchat, an ephemeral social media platform, phatic photo sharing is key to the success of the platform: Photo-sharing has been embedded in communicative media and supports the so-called "connected presence" (Rettberg, 2018; Niemelä-Nyrhinen and Seppänen, 2020). Niemelä-Nyrhinen and Seppänen (2020) argue that the omnipresence of the photo renders it "a kind of gesture suitable for opening and maintaining social exchange and bonds". In research on phatic visuals, a dichotomy is often drawn between traditional archival photographs — typically single images stored in photo albums — ephemeral contents, such as Instagram *stories* fade away (Niemelä-Nyrhinen and Seppänen, 2020; Rettberg, 2018). Another example of this is the young platform BeReal. On BeReal, ephemeral photos are shared directly with the community, without the availability of editing or curation of photos. The presented literature demonstrates how multifaceted the seemingly simple concept of phatic communication is.

3 Corpus compilation

With 37% of the population in Germany using Instagram (Newman et al., 2025), it is no surprise that Instagram reflects political moments of citizen engagement (Barbala, 2024). It is also a multimodal social network. As political messages are known to be communicated both in visuals and written language, it is especially important to study the platform (McNair, 2016). Politicians, for example, use Instagram visuals for strategic storytelling (Liebhart and Bernhardt, 2017). The Instagram caption is vital to this, as it is used to develop an engangig narrative (Towner and Muñoz, 2018).

We use an Instagram caption dataset structured by the German hashtags #WirSindMehr (we are more) and #NieWiederIstjetzt (never again is now). We build on work by Knierim and Heid (2025), who annotated the corpus for argument component detection. In their work, an argumentation model by Habernal and Gurevych (2017) was used. The model was developed and tested on user-generated web content, making it suitable for annotating social media discourse. The corpus includes both argumentative and non-argumentative sentences and comprises 13468 posts, with a sample of 1200 posts labeled. The corpus is suitable for the research question considered in our study, as it fits into the paradigm of political talk in social media. In the following paragraphs, we shortly introduce the discourses.

The corpus was collected via "crowdtangle"[1] and captures a timespan between 10/07/2023 and 03/31/2024. For this study, we only work with non-argumentative sentences from the corpus. This decision builds on the assumption that non-argumentative texts are more likely to be phatic. Like this, we investigate how many phatic messages we encounter within political discourse. The corpus contains 3013 non-argumentative sentences with 28432 tokens.

The upheavals against right-wing extremism (#WirSindMehr) and antisemitism (#NieWiederIstJetzt) provide a suitable dataset for investigating political communication. More than two million Germans protested using the slogans, while social media content on Instagram and Facebook was grammatized using the already mentioned hashtags. As hashtags enable and formalize actions and user activity, they can be used to trace user actions (Caliandro and Graham, 2020).

On January 10, 2023, the Correctiv research network uncovered a right-wing extremist meeting in Potsdam that called for the remigration of asylum seekers, foreigners with the right to stay in Germany and "non-assimilated citizens" in a secret plan (Correctiv 2024). Against the backdrop of the shift to the right in Germany and Europe in recent years, the unveiling of the secret meeting by the press had massive reverberations, with a total of two million people demonstrating in Germany in the following three months under the slogan #WirSindMehr.

Following the terrorist attack on Israel, the number of anti-semitic crimes committed in Germany increases by 240% by October 19, 2023. This leads to the formation of an alliance against antisemitism. On 09.11.2023 at the commemoration of the 85. anniversary of the november progroms, the words "Nie Wieder Ist Jetzt" are used for the first time. Eva Szepsi, a Holocaust survivor, uses the

---

[1] Crowdtangle was a legal way to extract data from Meta platforms until August 2024. However, the tool is no longer available. To extract data now, a highly formalised application process must be completed via the Meta API. The corpus can be released if legal review confirms its eligibility for public access.

phrase in the Bundestag on 31 January 2024 (Lelle, 2024).

In this paper, we investigate the following research questions: What is the amount of phatic communication in a dataset of non-argumentative sentences in German political posts? Which dimensions of the classes, phatic, mobilizing and informative, are most difficult to classify for LLMs? How do these difficulties compare between human annotation and LLM-based classification?

4 Annotation

4.1 Annotation guidelines

For the purpose of this study, non-argumentative texts were annotated and studied. From an inductive analysis of the texts, we derived three categories of non-argumentative sentences: Next to phatic texts, two other categories appear in the data: Sentences carrying information, such as reports from political events, or background information on persons or organizations. In addition, invitations or calls for engagement are common. This results in a three-fold classification of non-argumentative sentences, including phatic texts, informative texts, and mobilizing texts. Examples for all three categories are presented in Tab. 1. In this section, we outline the background and critera that informed our annotation guidelines, which served as the basis for the manual annotation process. The complete guidelines are displayed in the appendix (Tab. 6).

4.1.1 Phatic communication

Based on the literature review presented in section 2, we formulate the following guidelines for the phatic category:

- The text suggests a general sociability and affability without exchanging real information (Miller, 2017);
- The text evokes a sense of community (Miller, 2017) or a networked presence (Radovanovic and Ragnedda, 2012);
- The text evokes a sense of the feeling of a shared experience in real time (Rettberg, 2018);
- The text shares news to open a conversation (Duffy and Ling, 2020);
- The text can be uninteresting (Sarjanoja et al., 2013) or frivolous, boring and mundane (Rettberg, 2018).

4.1.2 Informative text

In the collection analyzed for this research, people often share information in their posts. While these sentences are non-argumentative and exist within a political context, they are neither phatic nor mobilizing.

- Background information on people, organizations, situations;
- Reports from political action or events;
- Announcements, specifications of time, place and date;
- Notes like: “No party symbols, but neutral banners are okay”;
- Short headlines.

4.1.3 Mobilizing text

More than two million people have protested under the slogans “Wir sind mehr” (“we are more”) and “Nie Wieder Ist Jetzt” (“Never again is now”). Part of this mobilization process is present in the nonargumentative posts of the corpus. In literature, specific concepts like the call-to-action are researched (Siskou et al., 2022; Knierim et al., 2024; Achmann-Denkler et al., 2024), we adhere to the more general notion of mobilization. The guidelines to annotate this category capture the following phenomena:

- Slogans, short invitations;
- Indirect calls, can be linked to a condition;
- Direct calls: “Give @mention a like and comment on their post to help the cause.”;
- Invitations such as: “All Magdeburgers are invited to participate in the commemoration of victims of antisemitism.”.

4.2 Annotation results

Two trained annotators achieve a kappa score of 0.62, suggesting moderate agreement. As is visible in Fig.1, the informative class is the dominant class for both annotators (annotator A:54.7%, annotator B:65.1%), while the phatic and the mobilizing class appear less often. The smallest differences becomes apparent in the mobilization class that annotator A identifies in 14.9% of the classes, while annotator B identifies it in 16% of the texts. The

Label	Example
Phatic	"You were great, you were colorful, and your voices are important!" "A great gesture from Wiesbaden's urban society against attempts at division" "The most beautiful picture of the week definitely comes from Cologne <3" "And we would like to thank all colleagues from our organization and beyond who are fighting with us for an open and diverse democracy."
Informative	"The *Eckernförde Round Table against the Right* will meet on 12.03.2024 - 7 p.m. in the AWO Family Centre & Citizens' Meeting Centre at the Rathausmarkt in Eckernförde." "There will be a rally and a vigil in Düsseldorf on Saturday and Sunday." "If you have flowers and grave candles, you are welcome to bring them, but it is not mandatory."
Mobilizing	"Come along and bring friends and family!" "Anyone who would like to make a contribution to the event is cordially invited to do so and should contact @mention (name@email.de) for coordination." "Show together with us: NEVER AGAIN IS NOW!"

Table 1: Examples for the phatic, informative and mobilizing class. (Translation from German by the authors.)

Possible labels	Example	Counts of disagreement
Mobilizing, Phatic	"So we'll see you tomorrow!"	86
Informative, Phatic	"On Sunday evening, around 6000 people lit up the banks of the Main during a chain of lights against antisemitism."	224
Mobilizing, Informative	"17th of February 2024, 17:00 on the market square in Dessau!"	94

Table 2: Ambiguous cases. (Translation from German by the authors.)

phatic class is identified in 30.5% of all texts (annotator 1) versus in 18.8% of all classes (annotator 2). Thus, the biggest discrepancy between the annotators lies in the phatic class. At the same time, we find that the biggest disagreement lies within the labeling of the phatic and the informative class (see Fig.1). Tab. 2 holds examples that are ambiguous.

5 Predicting phatic communication

5.1 Method

In order to explore the role of phatic communication in non-argumentative sentences quantitatively, we compare the performance of few-shot and zero-shot-settings of two large language models on detecting phatic communication. We are particularly interested in which dimensions of classes are most difficult to classify, and how these difficulties compare between manual annotation and LLM-based classification. We define the task as a multiclass classification, classifying the text as either phatic, informative, or mobilizing. Importantly, we investigate phatic texts in political communication, not in Instagram posts in general. The texts were preprocessed in the following way: errors from scraping were removed. We additionally removed hashtags and emojis.

For our classification, it is of specific interest to compare human annotations with generated annotations. Following Ziems et al. (2024), large language models have the potential to transform the pipeline in interdisciplinary research settings, if they are equipped in assisting with labeling tasks like classifications. In this setting, human annotators achieve only moderate agreement and see differing amounts of phatic and informative content, while agreeing on mobilizing content (see: Fig. 1). Therefore, we are interested in whether

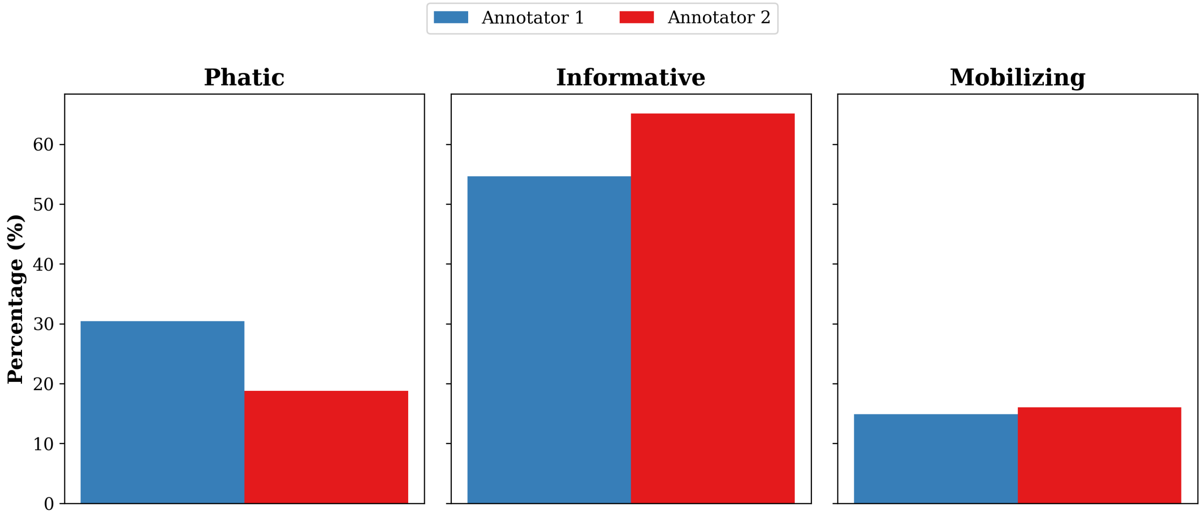


Figure 1: Distribution of informative, phatic and mobilizing sentences according to the annotations of annotator 1 and 2.

this ambiguity and uncertainty is reflected in model performance.

We test prompt drafts on every model and only present the best results. We follow the prompting guidelines provided by Ziems et al. (2024); Child et al. (2019) and give instructions after the context is provided. To reduce computational cost, we keep the prompt as short as possible with the GPT4o-model (temp=0), resulting prompts for both the zero- and fewshot settings that are different from those used with the open source model Mistral-7B-Instruct-v0.2 (Mistral) (Jiang et al., 2023). While we recognize that slight changes in prompt wording may affect the comparability of results across models, our goal is to showcase the best performance each model can achieve under optimal prompting conditions. Although Mistral had limited exposure to German during pretraining, we contrast its performance to that of GPT4o, as it is an open-source, mid-sized open-weight LLM.

5.2 Results

We present the results (Tab.: 3, 4, 5) for each class, comparing in turn both LLMs (few-shot (FS) vs. zero-shot (ZS) setting). In section 6, we compare the performance of the models with the human annotators.

For the classification of the informative class, both models perform well in the few-shot setting, while Mistral outperforms the GPT-4o model (f1 score = 0.87 vs f1 score = 0.79). The Mistral model shows no improvement when applied in a few-shot setting (f1 score = 0.87). GPT-4o improves in the few-shot-setting, not only for the informative class but in all classes. While the models showed satisfactory performance, both models struggle with lower recall, especially in zero-shot-settings. Compared to human annotators, we also see a difference within the identification of informative sentences, as annotator 2 is more inclusive with this category, identifying 1200 examples, while annotator 1 only identifies 1000 examples.

For the identification of the mobilizing class, GPT-4o performs best with an f1-score of 0.67 (see Tab. 5). For both zero-shot and few-shot settings, Mistral reaches high precision and lower recall, while GPT-4o has low precision but very high recall.

The models perform lowest for the detection of the phatic class, just like there is most ambiguity for the annotators in this class. As in the mobilization class, the GPT-4o few-shot setting (f1 score = 0.57) and the Mistral zero-shot setting yield the best results (f1 score = 0.58). Both models show higher precision and lower recall in both settings. The Mistral performs best in identifying the phatic class and the informative class, while GPT-4o performs best in identifying the mobilization class.

6 Error analysis

Both the manual annotation and the automatic classification show the difficulties in identifying phatic texts in our dataset of German political talk. This is consistent with Frenzel and Hautli-Janisz (2025) results who tested the idenitification of small talk on German verbatim transcribed Public Service Encounters (dataset: Espinoza et al. (2024)). While small talk is not identical to phatic political talk, a comparison is valid from a computational point of vie, as Espinoza et al. (2024) define small talk as polite conservation over light topics (Schneider, 1988) with the purpose of maintaining social relations for their study.

Both human annotators and models struggle most with identifying the phatic class (highest f1-score=0.58). While models perform best at identifying the informative class (highest f1-score=0.87), humans have the least dissent in annotating the mobilizing class. For the human annotators, some ambiguity exists between the mobilizing and the informative class (see: Tab. 2): While the example given informs about time and place of an event in Dessau, and could thus be labeled as informative, the exclamation mark makes the sentence sound like a call to join the event, and could thus be labeled as mobilizing class. Similarly, the informative and the phatic class can be confused when information is anecdotal or of little importance: Our example names the number of people, time and place, but focuses on the fact that lights lit up the banks of the Main river. While informative, the example could fit two aspects of phatic communication: On the one hand, it fits Miller (2017)'s definition of a general affability without exchanging real information, or, even more likely in this case, the text evokes a sense of community (Radovanovic and Ragnedda, 2012; Miller, 2017).

7 Discussion

In this paper, we aim at detecting phatic communication in political talk. For this purpose, a case study is conducted based on the hashtags of #NieWiederIstJetzt and #WirSindMehr. First, two annotators enriched the data with the labels phatic, mobilizing, and informative. Second, the performance of the classifiers Mistral and GPT-4o is tested, in both zero- and few-shot-settings. The annotators reach a kappa score of 0.62 which corresponds to moderate agreement. In general, the models perform better in a few-shot-setting. Only for classifying the informative class, there is no difference between the zero- and few-shot setting with Mistral- it yields the highest f1-score of 0.87 in both settings.

In summary, humans and models perform best in identifying the informative class, second best on the mobilizing class, and worst in identifying the phatic class. This also shows in the ambiguity between the annotators for the phatic class: Annotator 1 labels 12% texts more as phatic than annotator 2 (30.5% vs 18.8%).

It is well known that phatic communication is prevalent in social media. Some even argue that social media are phatic media in which communication without content has taken precedence (Miller, 2008). Nevertheless, researches have demonstrated the importance of phatic communication, as it fulfills a social function (Sarjanoja et al., 2013; Rettberg, 2018). While similar phenomena such as small talk have been studied from the lens of natural language processing (Frenzel and Hautli-Janisz, 2025) (also for 'candy speech', the shared task at GermEval 2025[2]). Others have studied the reception of phatic communication (Sarjanoja et al., 2013) or its different forms in various platform affordances (Niemelä-Nyrhinen and Seppänen, 2020; Radovanovic and Ragnedda, 2012; Rettberg, 2018).

In this paper, we ask how phatic political talk is in reality. Using non-argumentative sentences from a corpus of political talk from Instagram captions, we conclude the following: Firstly, annotators label between 18.8% and 30.5% of the corpus as phatic, which is a substantial amount. On the other hand, between 81.2% and 69.5% of the non-argumentative statements in the corpus is not phatic, but serves an informative or mobilizing function. Thirdly, the biggest dissent concerns the distinction between the phatic class and the informative class, demonstrating that the annotators struggle with the question whether a sentence is informative or not. Considering that the number of phatic texts in the argumentative part of the corpus (from where the non-argumentative sentences analyzed here have been gathered) is probably much lower, we conclude that political talk on Instagram is not primarily phatic. It rather has a high amount of informative content, and some mobilizing content.

Nevertheless, more work should be done to investigate phatic content in a quantitative manner: Our findings should be tested for political talk under other hashtags. Secondly, the distributions could be platform-specific, highlighting the importance to investigate distributions on Tiktok or other platforms. The analysis should be extended to other corpora and platforms. In addition, one

---

[2] yuliacl.github.io/GermEval2025-Flausch-Erkennung/

could investigate whether different political actors post different amounts of phatic communication. For example, it is likely that organizations post more informative and mobilizing content, while private users post phatically more often. It is also conceivable to perform a more fine-grained classification of phatic communication based on the codebook introduced in section 4.1.1. From a technological point of view, prompt engineering could also be applied to enhance the classification results. An idea could be to include stakes and bribes, as recommended by Pichler et al. (2025).

Limitations

This work is limited to the analysis of only one platform. This work could benefit from an additional annotation. Considering the identified sources of disagreement, the annotators would benefit from additional training.

Acknowledgments

We thank the annotators for their work. We thank the reviewers for their helpful comments.

	Mistral ZS	**Mistral FS**	**GPT4 ZS**	**GPT4 FS**
Precision	0.79	0.75	0.58	0.69
Recall	0.46	0.39	0.31	0.49
F1-Score	**0.58**	0.51	0.40	0.57

Table 3: Comparison of the different models for the classification of the phatic class.

	Mistral ZS	**Mistral FS**	**GPT4 ZS**	**GPT4 FS**
Precision	0.85	0.82	0.87	0.81
Recall	0.89	0.92	0.67	0.78
F1-Score	**0.87**	**0.87**	0.76	0.79

Table 4: Comparison of the different models for the classification of the informative class.

	Mistral ZS	**Mistral FS**	**GPT4 ZS**	**GPT4 FS**
Precision	0.59	0.67	0.36	0.53
Recall	0.62	0.49	0.90	0.92
F1-Score	0.60	0.57	0.51	**0.67**

Table 5: Comparison of the different models for the classification of the mobilization class.

References

Michael Achmann-Denkler, Jakob Fehle, Mario Haim, and Christian Wolff. 2024. Detecting calls to action in multimodal content: Analysis of the 2021 german federal election campaign on instagram. *arXiv preprint arXiv:2409.02690*.

Astri Moksnes Barbala. 2024. Reassembling# metoo: Tracing the techno-affective agency of the feminist instagram influencer. *Convergence*, 30(3):992–1007.

Shelley Boulianne. 2020. Twenty years of digital media effects on civic and political participation. *Communication research*, 47(7):947–966.

Alessandro Caliandro and James Graham. 2020. Studying instagram beyond selfies. *Social Media + Society*, 6(2):2056305120924779.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Andrew Duffy and Rich Ling. 2020. The gift of news: Phatic news sharing on social media for social cohesion. *Journalism studies*, 21(1):72–87.

Ingrid Espinoza, Steffen Frenzel, Laurin Friedrich, Wassiliki Siskou, Steffen Eckhard, and Annette Hautli-Janisz. 2024. PSE v1.0: The first open access corpus of public service encounters. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13315–13320, Torino, Italia. ELRA and ICCL.

Steffen Frenzel and Annette Hautli-Janisz. 2025. Identifying small talk in natural conversations. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 272–277, Albuquerque, New Mexico. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational linguistics*, 43(1):125–179.

Roman Jakobson. 1968. Poetry of grammar and grammar of poetry. *Verbal Art, Verbal Sign, Verbal Time*, 37:604.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Aenne Knierim, Michael Achmann-Denkler, Ulrich Heid, and Christian Wolff. 2024. Divergent discourses: A comparative examination of blackout Tuesday and #BlackLivesMatter on Instagram. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 451–458, Pisa, Italy. CEUR Workshop Proceedings.

Aenne Knierim and Ulrich Heid. 2025. Argumentation in political empowerment on Instagram. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 97–108, Albuquerque, New Mexico. Association for Computational Linguistics.

Aenne C. K. Knierim, Ulrich Heid, and Thomas Mandl. 2025. Politisches Empowerment in den sozialen Medien: Ein Ansatz zur sprachlichen Charakterisierung am Beispiel von #WirSindMehr. In *Datenströme und Kulturoasen — Die Informationswissenschaft als Bindeglied zwischen den Informationswelten. 18. Internationales Symposium für Informationswissenschaft (ISI 2025)*, pages 296–314, Chemnitz, Germany.

Nikolas Lelle. 2024. Demos gegen rechts: Was heißt Nie wieder? — taz.de. `https://taz.de/Demos-gegen-rechts/!5987599/`. [Accessed 12-08-2025].

Karin Liebhart and Petra Bernhardt. 2017. Political storytelling on instagram: Key aspects of alexander van der bellen's successful 2016 presidential election campaign. *Media and Communication*, 5(4):15–25.

Bronislaw Malinowski. 1967. The problem of meaning in primitive languages. *Meaning in Meaning*.

Brian McNair. 2016. Image, political. In R C M Maia, K Ikeda, H Wessler, G Mazzoleni, and K G Barnhurst, editors, *The international encyclopedia of political communication, Volume 1*, pages 505–513. Wiley-Blackwell, United States of America.

Vincent Miller. 2008. New media, networking and phatic culture. *Convergence*, 14(4):387–400.

Vincent Miller. 2017. Phatic culture and the status quo: Reconsidering the purpose of social media activism. *Convergence*, 23(3):251–269.

Nic Newman, A Ross Arguedas, Craig T Robertson, Rasmus Kleis Nielsen, and Richard Fletcher. 2025. *Digital news report 2025*. Reuters Institute for the study of Journalism.

Jenni Niemelä-Nyrhinen and Janne Seppänen. 2020. Visual communion: The photographic image as phatic communication. *New Media Soc.*, 22(6).

Axel Pichler, Janis Pagel, and Nils Reiter. 2025. Evaluating LLM-Prompting for Sequence Labeling Tasks in Computational Literary Studies. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 32–46.

Danica Radovanovic and Massimo Ragnedda. 2012. Small talk in the digital age: Making sense of phatic posts. In *Proceedings of the WWW'12 Workshop on 'Making Sense of Microposts', Lyon, France, April 16, 2012*, volume 838 of *CEUR Workshop Proceedings*, pages 10–13. CEUR-WS.org.

Jill Walker Rettberg. 2018. Snapchat: Phatic communication and ephemeral social media. *Appified: culture in the age of apps*, pages 188–96.

Ari-Heikki Sarjanoja, Minna Isomursu, and Jonna Häkkilä. 2013. Small talk with facebook: Phatic communication in social media. In *Proceedings of International Conference on Making Sense of Converging Media*, pages 118–121.

Klaus P Schneider. 1988. Small talk: Analysing phatic discourse. *(No Title)*.

Wassiliki Siskou, Clara Giralt Mirón, Sarah Molina Raith, and Miriam Butt. 2022. Automatized detection and annotation for calls to action in latin-american social media postings. In *The 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2022)*, page 65.

Terri L. Towner and Caroline Lego Muñoz. 2018. Picture perfect? the role of instagram in issue agenda setting during the 2016 presidential primary campaign. *Social Science Computer Review*, 36(4):484–499.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Appendix: Annotation guidelines

Table 6: Annotation instructions for labeling non-argumentative sentences as phatic, informative, or mobilizing.

Class	Guidelines
Phatic communication	• If the text conveys sociability or affability without transmitting substantive information • If the text that evokes a sense of community • If the text creates the impression of sharing an experience in real time • If the text shares news as a conversation starter • If the text appears uninteresting, boring, or mundane
Informative text	• If the text contains background information about people, organizations, or situations • If the text include reports on political events or actions • If the text is an announcement sharing time, place, or date details • If the text include specific notes, e.g., “No party symbols, but neutral banners are okay”. • If the text marks a short headline
Mobilizing text	• If the text is a slogan or short invitations aimed at motivating action • If the text contains a direct calls, e.g., “Give @mention a like and comment on their post to help the cause.” • Include indirect calls to action • If there is an invitations to participate in events, e.g., “All Magdeburgers are invited to participate in the commemoration of victims of antisemitism.”

TRAVELWARN-Crawler: Constructing longitudinal datasets of government-issued travel warnings for political and social science research

Laura Braun
Center for Crisis Early Warning
University of the Bundeswehr Munich
Neubiberg, Germany
`laura.braun@unibw.de`

Christian Oswald
Center for Crisis Early Warning
University of the Bundeswehr Munich
Neubiberg, Germany
`christian.oswald@unibw.de`

Abstract

Historical travel warnings and advisories offer a record of how governments perceive and communicate country-specific risks over time, yet large-scale quantitative analyses have been rare due to missing, fragmented, and frequently overwritten web data. We present *TRAVELWARN-Crawler*, an open pipeline that collects, recovers, cleans, stores, and standardizes advisories and warnings from archived web sources. Using the Internet Archive's Wayback Machine and issuer-specific Scrapy spiders, we reconstruct up to three decades of timelines for the United States (US), the United Kingdom (UK), and Australia. The data can be used to support comparative research in political communication, international relations, and tourism studies. Descriptively, the UK issues more country pages and updates than the US and Australia. Cross-issuer agreement about severity is modest with the highest pairwise level agreement for US–Australia (∼49%). These results indicate substantial heterogeneity in how close allies communicate travel risk and underscore the value of reproducible, textual data beyond numeric severity levels alone.

1 Introduction and Motivation

The U.S. Department of State issued what is widely considered the first modern travel warning when it cautioned its nationals at the outbreak of World War I in 1914 (Löwenheim, 2007). Since the late 1990s, such warnings have increasingly become accessible online and on a regular basis, shaping individual travel decisions and redirecting tourism revenue. Beyond their immediate practical value, they also signal a government's assessment of specific countries. Although they are potentially important, researchers lack comprehensive longitudinal data. Each foreign ministry maintains its own website and frequently overwrite pages in place. There is no common official API for retrieving archived warning texts. The paper demonstrates how to navigate this terrain. We reconstruct advisory timelines for three issuers (US, UK, Australia) by requesting archived captures via the Wayback Machine API, extracting text content and metadata, and processing and standardizing the results to store in a database, to provide a ready-made and easily downloadable dataset.

Government-issued travel advisories represent a unique combination of time-connected threat assessment and official foreign policy signaling. Each text is a judgment of how the issuing state evaluates security, health, and political conditions abroad at a given moment. Since ministries publish these assessments under their own seal, language and severity levels may encode far more than objective risk. Having such textual data allows to investigate whether they also reflect strategic calculations about alliance obligations, latent disputes, or reputational costs (Chu et al., 2021; Babey, 2019; Kebede, 2018; Sharpley et al., 1996). We harvest these text data at scale and reconstruct their full issuing history to create the first longitudinal warning corpus that scholars can reuse and provide a step-by-step guide for extending the dataset to additional issuers. We focus on the US, UK, and Australia as a proof of concept and a foundation for comparative analysis. The countries are three founding members of the Five Eyes intelligence alliance, and we explore how their informal cooperation is reflected in travel warning decisions. Although we start with English sources, recent advances in automated translation may enable the inclusion of other languages.

2 Implementation

2.1 Data access

We use the Internet Archive's Wayback Machine (IAWM) API to retrieve historical digital artifacts of travel warnings and advisories from archived snapshots of provider webpages. The IAWM pre-

Proceedings of the 5th Workshop on Computational Linguistics for the Political and Social Sciences (CPSS), pages 23–29
September 10-11, 2025 ©2025 Association for Computational Linguistics

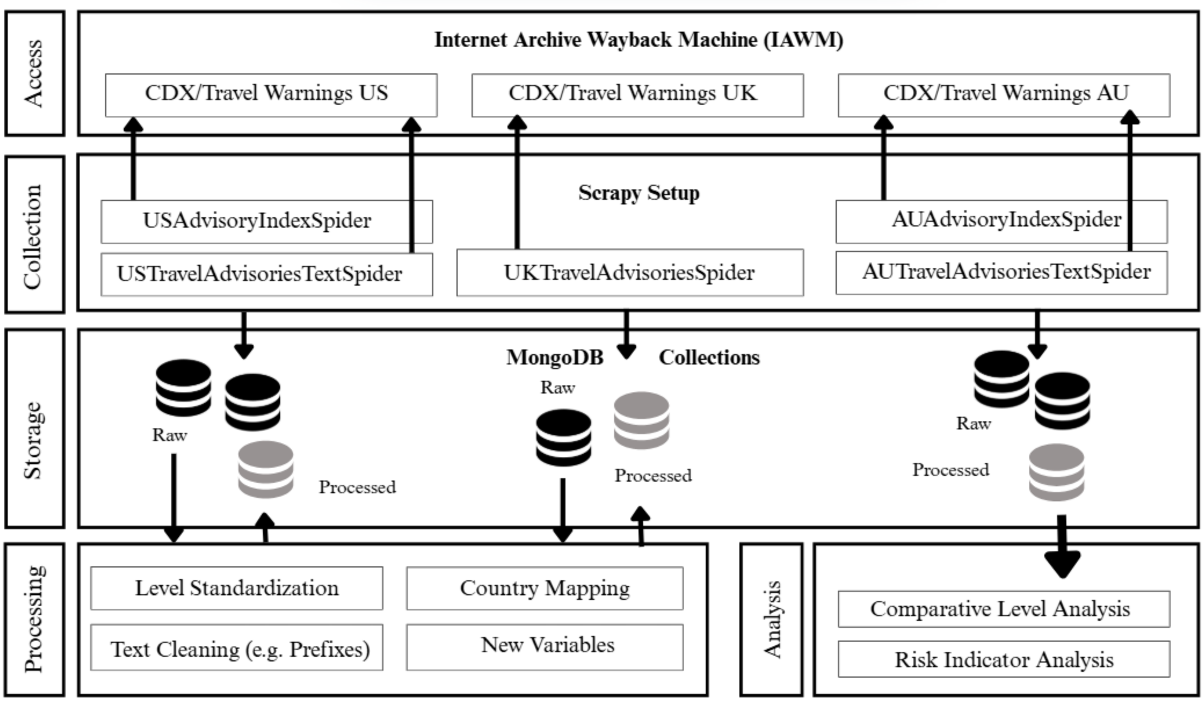


Figure 1: Implementation of dataset generation — Access, Collection, Storage

serves time-stamped URLs since 1996 and indexes more than 800 billion pages.[1] Its archival record turns otherwise ephemeral content into a longitudinal resource for scholars. A validation study demonstrates that the IAWM website age and update count measures are reliable, as the authors find convergent validity with domain-registry dates, confirming the archive's suitability as a research resource (Murphy et al., 2007). Since the foreign ministry websites we analyze went online in 1996 or 1997 (Löwenheim, 2007), the IAWM offers a long archive history with sparse coverage until the mid-2000s and increasingly dense coverage thereafter.

Previous work has outlined a complete social science workflow for sampling, crawl design, variable extraction, and transformation into structured data (Arora et al., 2016). Our approach differs in retrieving the complete archive for each issuer, prioritizing depth over breadth. This introduces practical challenges such as changing URLs, layouts, and HTML structures. We apply custom crawler logic that uses issuer- and time-specific URL prefixes to query the IAWM's CDX endpoint. Other archives are paid services or lack comparable coverage. Some issuers provide partial official archives (e.g., OSAC.gov for U.S. alerts since 2004, the UK Government archive since 2013), but these are either restricted for scientific use due to licensing or offer fewer snapshots. The IAWM with its CDX API is thus the best choice for standardized, near-complete timelines from most issuers, without copyright restrictions.[2]

2.2 Data collection and storage

We implement a modular, automated pipeline in `Scrapy`. Scrapy's abstractions (Spiders, Items, Pipelines) separate crawling from extraction and storage, while built-in concurrency, retry logic, and scheduling enable large-scale, polite crawling of mostly static content, lighter than headless-browser approaches. For the US and Australia, a two-step spider design first queries the CDX API for *one* index-page snapshot per day, extracting country links, update dates, and any listed metadata (e.g. level). For each new (country, update-date) pair, we then fetch the first available archived copy of the corresponding country page to extract full text and metadata. This design (i) detects missed updates even when country pages are sparsely archived, and (ii) avoids unnecessary downloads. For the UK, whose index pages lack explicit update dates, we instead check country pages daily and re-scrape only when their on-page update date changes. We

---

[1] `https://archive.org/about/`

[2] IAWM licensing is generally permissive, but site-specific terms should always be verified, although government pages are typically open-licensed.

observed that UK country pages also contain rich, multi-section content (safety/security, health, local laws, etc.), which can trigger additional updates and thus increase revision frequency.

All advisories are stored in MongoDB, one collection per issuer. Each document includes at least `country`, `warning_date`, the IAWM `timestamp_snapshot`, the `source_url`, and the extracted `advisory_text`. We upsert on a compound key (country, update date) to prevent duplicates and preserve a stable pointer back to the archived source for full traceability.

2.3 Data processing

After collection, we standardize the records to allow cross-issuer comparisons. First, we normalize country names to a canonical mapping to accommodate spelling variants and historical names. Second, we harmonize issuer-specific severity descriptions to a unified four-tier scale using regular expressions, explicitly accounting for policy changes over time (e.g. pre-2018 U.S. materials without formal levels; UK advisory phrasing without Level 1/2 labels). We retain the original text for full transparency. Although regex-based mapping is transparent and reproducible, it can be brittle in the presence of negation or nuanced phrasing (e.g. "no longer advised not to travel…"). In practice, a small residual remains: for Australia, n=158 updates could not be mapped; for the United States, n=23 entries are labeled *Other*; and for the United Kingdom, because Level 3 and Level 4 are explicitly indicated, cases ambiguous between Level 1 and Level 2 default to Level 2. In future work, we will evaluate supervised models (e.g. ConfliBERT; Brandt et al., 2024) and LLM-assisted level assessment to reduce regex-based classification noise.

3 Results

We reconstruct longitudinal advisory timelines for the United States, the United Kingdom, and Australia and harmonize them to a common four–level scale. Coverage is complete for the US and Australia from 1997/1998 onward and extends back to 2013 for the UK. Descriptively, issuers differ in both scope and frequency.

We restrict descriptive comparisons to the overlapping ten–year window (2014–2024) in which all three issuers are covered. Table 1 summarizes this period. The UK issues the most updates (15,207) across the largest destination set (225 unique countries), followed by Australia (6,530 with 181 unique countries) and the US (3,101 with 210 unique countries). Consistent with the editorial breadth of UK pages, only ≈4% of UK updates coincide with a level change (611/15,207), compared to ≈8.4% for Australia (547/6,530) and ≈31.9% for the US (990/3,101). In other words, the typical US update is more likely to reflect a substantive change in severity, whereas UK updates more often revise text within informational sections. UK pages also yield a larger set of destinations that remain at Level 4 throughout the period ("constant L4"): 33 for the UK versus 11 for Australia and 5 for the US, reflecting both coverage scope and differing issuer thresholds to maintain a sustained "Do not travel" classification.

Figure 2 tracks the monthly average advisory level for each issuer, smoothing with a 3-month moving average to reduce month-to-month noise. The dashed line marks the US shift to the four-tier system in 2018; the gray band marks the global COVID shock. Two patterns stand out. First, during COVID all three issuers converge upward, indicating a sharp, broad-based tightening that later relaxes at different speeds. Second, outside the COVID window, UK and Australia trend closely together on average, while the US is modestly higher in the pre-2018 period and again during the 2021–2022 normalization phase.

Figure 3 provides exemplary trajectories. For **Ukraine**, Australian levels move from Level 1–2 in the early 2000s to pronounced step-ups around 2014 and again from 2022 onward, while the US remains more persistently at Level 3–4. For **Lebanon**, both issuers spend extended periods at Level 3–4, punctuated by frequent revisions (black lines) that do not always change the posted level. For **South Korea**, Australia stays mostly at Level 1 with occasional upticks, whereas the US displays the system-wide pandemic jump to Level 4 and subsequent relaxation. The dense vertical markers illustrate that issuers revise advice frequently, even when the numeric level remains, thereby underscoring the value of archiving text alongside levels.

To quantify convergence, we align updates within countries using a nearest-neighbor match inside a ±15-day window (see Appendix for details) and then compare levels. Across all matched updates, **US–Australia** aligns the most, agreeing on **48.5%** of 890 matches (432 agreements). **UK–US** agrees on **33.1%** of 2,533 matches (839), and **UK–Australia** on **29.3%** of 5,299

Table 1: Summary statistics (2014–2024) for UK, US, and Australia advisories.

Issuer	L3 count	L4 count	Unique countries	Level changes	Updates	Most changes	Constant L4
UK	2,810	3,904	225	611	15,207	China (41)	33
Australia	651	2,014	181	547	6,530	Bangladesh (8)	11
US	844	758	210	990	3,101	D.R. Congo (21)	5

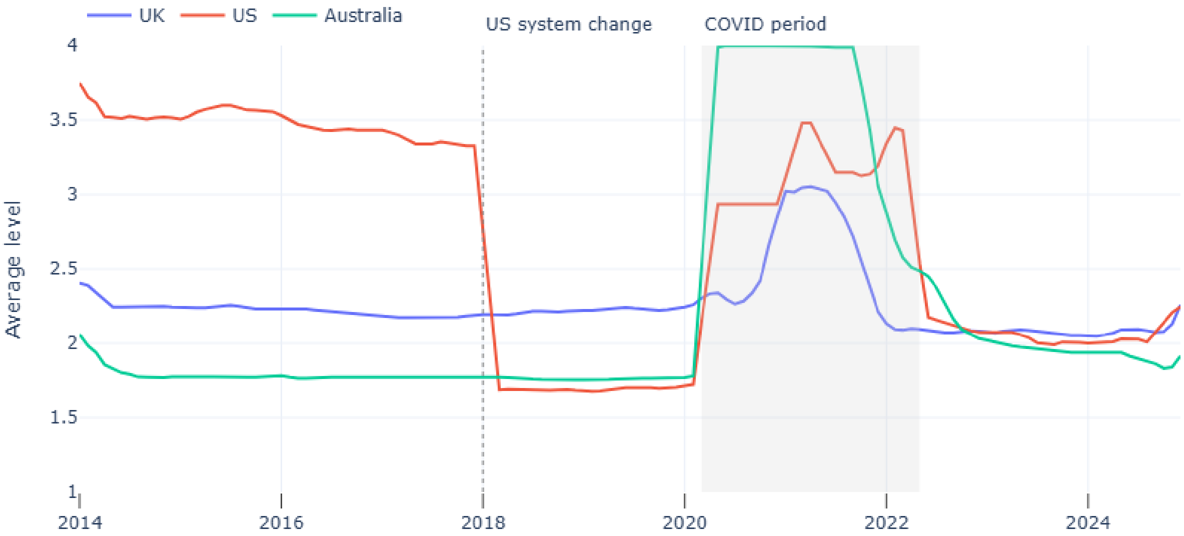


Figure 2: Monthly average advisory level across a common basket of countries for the UK, US, and Australia. Lines show 3-month moving averages; shaded bands are interquartile ranges. The dashed vertical line marks the US system change in 2018; the light gray rectangle marks the COVID period (2020–2022).

(1,550). When restricting to events where any issuer assigns Level 3/4, agreement rates drop: **37.5%** for US–Australia (238/635), **29.4%** for UK–US (584/1,987), and **23.3%** for UK–Australia (852/3,662). Requiring *all three* issuers to coincide produces very low agreement: **16.9%** across all matched triples (203/1,199) and **15.9%** for the Level 3/4 subset (165/1,039). Substantively, even among close allies, severity judgments are far from harmonized. The comparatively higher US–Australia alignment is consistent with shared terminology and, post-2018, more comparable US level definitions; by contrast, the UK's revision-intensive advice style lowers the probability of within-window convergence at the same numeric level.

The summary counts also reveal distinct issuer profiles. The UK posts many more absolute Level 3 and Level 4 observations (2,810 and 3,904), but as a *share* of all UK updates, these are smaller than Australia's, which has fewer updates overall, yet a higher fraction of high-severity postings. The US sits between the UK and Australia in scope, but shows the highest ratio of level-changing to total updates, reflecting more tightly scoped, advisory-centric editing pre-2018. Importantly, the aggregate lines in Figure 2 are computed on the *common basket* of countries to minimize composition bias.

Low pairwise and triple alignment can arise from (i) different thresholds for moving between Level 3 and Level 4, (ii) non-synchronous response timing around the same underlying event, and (iii) composition effects (issuers emphasizing different destination portfolios and sub-national guidance). Figure 3 shows that many micro-revisions do not alter levels, while the COVID panel in Figure 2 illustrates the opposite case, globally synchronized shocks that temporarily compress issuer differences and then re-diverge as governments normalize at different speeds.

A large share of updates, especially for the UK and Australia, revise or expand guidance without changing the numeric level. Most are minor editorial adjustments (wording, formatting, refreshed links), but many introduce substantively useful text: added/removed specific risk indicators, tighter regional exclusions, updated entry/exit rules, or clarifications of recent incidents. These "text-only" re-

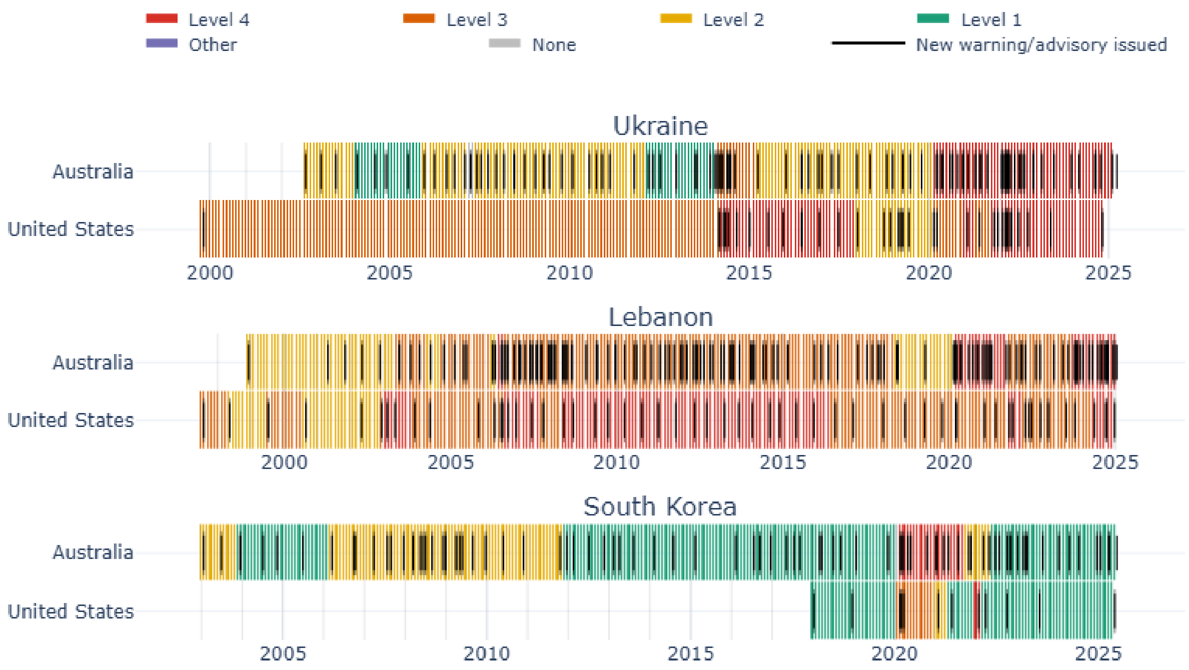


Figure 3: Monthly travel-advisory levels for three illustrative cases (Ukraine, Lebanon, South Korea). Rows show issuer (Australia, United States). Black marks indicate publication days of new warnings/advisories.

visions still signal issuer attention and policy stance even when the severity is unchanged.

4 Discussion

Our results show that even among close allies, advisory severity is far from harmonized and that issuers differ markedly in update cadence. A natural next step is to exploit the text itself rather than levels alone. Information extraction and named-entity recognition (NER) can recover subnational geography (regions, corridors, border zones) and relevant actors, enabling granular, region-level timelines layered onto the national series. Topic discovery and topic classification can map risk drivers (crime, terrorism, civil unrest, health, disasters, kidnapping, wrongful detention) and track how their composition shifts across issuers and time; change-point detection on topic proportions or keyphrase series can sharpen the timing of substantive updates that occur without a level change. Cross-issuer harmonization can be improved with supervised classifiers and LLM-assisted labeling to normalize historical phrasing into a shared taxonomy of risks and severities. Together, these textual transformations open the door to event-aware models of advisory dynamics that incorporate both global shocks and country-specific triggers, and to theory tests on alliance behavior, responsiveness, and the strategic use of language in official risk communication.

Limitations

Archive coverage. Our timelines are bounded by what the Internet Archive preserved. Gaps exist, particularly in the late 1990s and for low-traffic URLs. Missing snapshots may selectively underrepresent short-lived advisories.

Site drift and parsing. Ministry redesigns and evolving HTML lead to fragile extraction rules. Although our Scrapy spiders handle many variants, long-term maintenance is required. Edge cases (e.g. split country pages, temporary microsites) can still slip through.

Harmonization choices. Mapping issuer-specific phrasing to a four-level scale inevitably introduces noise (e.g. negations in "no longer advised to…"). Our regex strategy is transparent and reproducible but imperfect; supervised or LLM-assisted classifiers are a promising replacement once labeled data exist.

Structural changes. The US system changed in 2018, and pandemic-era global notices created atypical level spikes. Such policy regime shifts should be modeled explicitly when drawing causal inferences from the time series.

Generalizability. We focused on three English-language issuers for proof of concept. Extending to non-English issuers will require additional country-name normalization and machine translation, and coverage may vary by local archival practices.

Ethical considerations

The data processed and analysed through this pipeline consists solely of publicly available travel advisories from official government sources (United States, United Kingdom, and Australia). The dataset contains no personally identifiable information (PII) or other sensitive data.

Acknowledgments

We thank David Bencek, André Bluhm, Vanessa Gottwick, Marius Hofmann and Daniel Racek, as well as the editors and three anonymous reviewers, for their helpful comments and feedback. The Center for Crisis Early Warning (Kompetenzzentrum Krisenfrüherkennung) is funded by the German Federal Ministry of Defense and the German Federal Foreign Office. The views and opinions expressed in this article are those of the author(s) and do not necessarily reflect the official policy or position of any agency of the German government.

Data and code availability

Upon publication, Scrapy spiders, processing scripts, and analysis notebooks will be released, together with issuer-specific schemas and harmonization code, under an open license. In subsequent work, we will publish the full dataset accompanied by an in-depth analysis paper that documents construction choices, validation checks, and known limitations.

References

Sanjay K. Arora, Yin Li, Jan Youtie, and Philip Shapira. 2016. Using the wayback machine to mine websites in the social sciences: A methodological resource. *Journal of the Association for Information Science and Technology*, 67(8):1904–1915.

Nicholas George Babey. 2019. The Politics of Travel Advisories: Foreign Policy and Error in Canada's Traveller Information Program. *The Journal of Intelligence, Conflict, and Warfare*, 2(1):15–33. Number: 1.

Patrick T. Brandt, Sultan Alsarra, Vito J. D'Orazio, Dagmar Heintze, Latifur Khan, Shreyas Meher, Javier Osorio, and Marcus Sianan. 2024. ConfliBERT: A Language Model for Political Conflict. *arXiv preprint*. ArXiv:2412.15060 [cs].

Yinxiao Chu, Xiaoyu Huang, and Tao Jin. 2021. Political relations and tourism: evidence from China. *Applied Economics*, 53(45):5281–5302. Publisher: Routledge _eprint: https://doi.org/10.1080/00036846.2021.1922591.

Nafbek Solomon Kebede. 2018. The Fate of Tourism during and in the Aftermath of Political Instability: Ethiopia Tourism in Focus. *Journal of Tourism & Hospitality*, 7(1):1–7. Publisher: Longdom Publishing S.L.

Oded Löwenheim. 2007. The Responsibility to Responsibilize: Foreign Offices and the Issuing of Travel Warnings. *International Political Sociology*, 1(3):203–221.

Jamie Murphy, Noor Hazarina Hashim, and Peter O'Connor. 2007. Take Me Back: Validating the Wayback Machine. *Journal of Computer-Mediated Communication*, 13(1):60–75.

Richard Sharpley, Julia Sharpley, and John Adams. 1996. Travel advice or trade embargo? The impacts and implications of official travel advice. *Tourism Management*, 17(1):1–7.

A Appendix

Collections, document examples and structure: For the US and Australia, each advisory is represented by two related documents:
(1) an "index" record (e.g. `us_advisories_index`) containing update metadata, and
(2) a "full-text" record (e.g. `us_advisories`) containing the parsed text and additional metadata (See examples below). For the UK, the index collection is not available; only full-text records exist.

Australia example (Lebanon, 1998):

```
{
  "country": "Lebanon",
  "warning_date": "1998-12-08",
  "advisory_text": "Australians
      travelling or resident in Lebanon
      should keep [...],
  "date_updated": "1998-12-08",
  "source_url": "https://web.archive.org
      /web/19990203024850/http:/www.dfat
      .gov.au/consular/advice/lebanon.
      html",
  "timestamp_snapshot": "19990203024850"
}
```

UK example (Lebanon, 2013):

```
{
  "country": "Lebanon",
  "level": "against all travel",
  "warning_date": "2013-03-28",
  "advisory_text": "Summary Still
      current at: 7 April 2013 [...]",
  "source_url": "https://web.archive.org
      /web/20130407005707/https://www.
      gov.uk/foreign-travel-advice/
      lebanon",
  "timestamp_snapshot": "20130407005707"
}
```

US example (Lebanon, 2024):

```
{
  "country": "Lebanon",
  "warning_date": "2024-12-27",
  "level": "Level 4: Do Not Travel",
  "advisory_text": "Updated to reflect
      the lifting of ordered departure
      [...]",
  "tooltips": [
    "Other: There are potential risks
       not covered [...]",
    "Kidnapping/Hostage Taking: [...]",
    "Civil Unrest: [...]",
    "Terrorism: [...]",
    "Crime: [...]"
  ],
  "source_url": "https://web.archive.org
      /web/20250103135018/https://travel
      .state.gov/content/travel/en/
      traveladvisories/traveladvisories/
      lebanon-travel-advisory.html",
  "timestamp_snapshot": "20250103135018"
}
```

Appendix A: Cross-issuer matching and agreement

We quantify cross-issuer alignment by per-country nearest-neighbor matching. For a given issuer A with update dates $t^{A}_{c,i}$ for country c, and issuer B with dates $t^{B}_{c,j}$, we match each $t^{A}_{c,i}$ to the chronologically nearest $t^{B}_{c,j}$ within a symmetric tolerance window of ± 15 days. Ties are broken by absolute time distance; updates without a counterpart inside the window are dropped. Agreement is computed over matched pairs (or triples) by comparing harmonized levels.

Table 2: Pairwise and triple agreement of advisory *levels*. Rows matched by country and nearest update within ± 15 days. "All" uses all matched updates; "L3/L4" restricts to matches where any issuer assigns Level 3 or Level 4.

Pair	Rows (all)	Agree % (all)	Agree n (all)	Rows (L3/L4)	Agree % (L3/L4)	Agree n (L3/L4)
UK–Australia	5,299	29.25	1,550	3,662	23.27	852
UK–US	2,533	33.12	839	1,987	29.39	584
US–Australia	890	48.54	432	635	37.48	238
UK–US–Australia	1,199	16.93	203	1,039	15.88	165

Causally fooled in the name of being honest? Evaluating causal extraction in LLMs for political text

Paulina García Corral
Hertie School
`corral@hertie-school.org`

Hannah Béchara
Hertie School
`bechara@hertie-school.org`

Krishnamoorthy Manohara
Hertie School
`manohara@hertie-school.org`

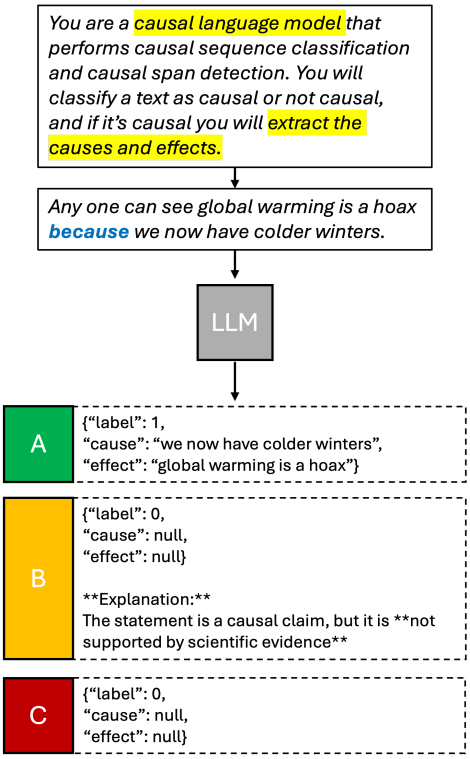
Slava Jankin
University of Birmingham
`v.jankin@bham.ac.uk`

Abstract

Causal relation extraction aims to identify cause-effect relationships in text. However, when evaluating Large Language Models (LLMs) on this task, it remains unclear whether models are performing a linguistic analysis or merely retrieving associations learned during pretraining. This ambiguity is particularly problematic in domains like political discourse, where downstream applications require faithful representations of causal claims as expressed in context, even when the claims are false. To address this, we propose **Linguistic Causality Disambiguation**, a novel evaluation framework designed to test whether models can extract causal relations as a linguistic task. Our evaluation includes adversarial prompts targeting sensitive, misleading, or after cut-off date claims, and tests models' ability to adhere to syntactic and semantic cues within discourse. Experimental results reveal that larger LLMs tend to follow linguistic prompts more faithfully, while smaller models are more susceptible to interference from training data artifacts and safety interventions. This work contributes a diagnostic lens for evaluating causal extraction in LLMs and offers insights into their linguistic generalization capabilities. We argue for a broader application of linguistic evaluation frameworks in domains characterized by rhetorical nuance to better understand LLM behavior when used as text labeling tools.

1 Introduction

Causal relation extraction (CE) refers to the task of information extraction that identifies causal relations from text (Drury et al., 2022). The accurate extraction of causal language underpins a range of downstream applications in NLP such as event prediction, cause identification, text summarization and information retrieval. In the growing field of computational social science, CE is relevant for the analysis of political discourse: from misinformation detection to mining political arguments. With the growing reliance on LLMs for traditional computational linguistic tasks, it's still unclear how these models perform CE, and if using LLMs for annotating causal structures of claims that are false produces accurate representations of the claim.



You are a causal language model that performs causal sequence classification and causal span detection. You will classify a text as causal or not causal, and if it's causal you will extract the causes and effects.

Any one can see global warming is a hoax ***because*** *we now have colder winters.*

{"label": 1,
"cause": "we now have colder winters",
"effect": "global warming is a hoax"}

{"label": 0,
"cause": null,
"effect": null}

Explanation:
The statement is a causal claim, but it is **not supported by scientific evidence**

{"label": 0,
"cause": null,
"effect": null}

Figure 1: We test to see how LLMs perform causal extraction. We present three potential responses: in A) the model parses the text linguistically, in B) the model produces a null output and explains that it recognizes the linguistic causal claim. Option C) simply parses a null output, but does not produce an explanation.

> **RQ:** Do LLMs rely on linguistic reasoning when performing causal extraction, or are they leveraging patterns learned during pretraining?

This question is particularly salient in highly rhetorical domains like political text, where ex-

Proceedings of the 5th Workshop on Computational Linguistics for the Political and Social Sciences (CPSS), pages 30–47
September 10-11, 2025 ©2025 Association for Computational Linguistics

pressions of causality are often exaggerated or factually incorrect. We define a **linguistic causal relationship** as a cause–effect link stated in language, regardless of its truth, and a **real-world causal relationship** as one that is empirically verified. For example, "Climate change is not real because there's an increase in cold snaps" expresses a linguistic causal relationship (cold snaps $\rightarrow$ no climate change) that is scientifically false. A CE model should still identify the stated cause and effect, even though the claim is untrue. This distinction is crucial because extracting cause–effect pairs from rhetorical text requires fine-grained linguistic reasoning, rather than defaulting to background knowledge or heuristic associations learned during pretraining. Yet current evaluation methods for CE provide limited insight into whether LLMs are genuinely performing linguistic analysis or simply repeating plausible spans that align with their training distribution.

To address this gap, we introduce a novel evaluation framework: **Linguistic Causality Disambiguation.** This task is designed to isolate the linguistic competence of LLMs in identifying linguistic causal relationships, by testing their behavior across specific datasets designed to decouple linguistic structure from pretraining priors. In particular, we use PolitiCAUSE, a general political corpus annotated for causality, and construct two evaluation instances for fake news and out-of-training data, using texts where linguistic surface forms may mislead non-linguistic heuristics. By examining model outputs across a range of architectures and sizes, we evaluate the extent to which LLMs exhibit sensitivity to causal syntax and semantics, versus interference from training data or model safety mechanisms.

Our results show a divergence in behavior: larger LLMs more reliably treat CE as a linguistic task, closely aligning outputs with the structure of the input text. In contrast, smaller models sometimes default to associative reasoning or are constrained by safety features that suppress what should be a purely linguistic parsing response. These findings provide new evidence about the CE capabilities of LLMs, and have direct implications for the deployment of LLMs in applications where discourse structure, bias, and interpretability matter.

In this paper, we make the following contributions: 1) We broadly test LLM usefulness for CE by evaluating models of varying sizes and configurations using a general political corpus annotated for causality. 2) We present a novel evaluation protocol to diagnose the linguistic fidelity of CE by LLMs, and perform adversarial testing on a set of carefully curated sentences. 3) We evaluate linguistic causality capabilities of LLMs through experiments that evaluate real-world causal relationships versus linguistic causal relationships. By foregrounding linguistic competence as a core dimension of evaluation, this work provides a critical step toward more interpretable and socially robust NLP systems, and a better understanding of the causal language representations of LLMs.

2 Related work

2.1 Causal Relation Extraction

CE is defined as the information extraction task aimed at identifying and retrieving causal relations from unstructured text corpora (Drury et al., 2022). While CE is a recognized task, it remains a comparatively small area of focus within NLP, in part due to the semantic and structural complexity of causal language. While causality can be expressed through explicit discourse markers, such as "because", and "therefore", it can also be expressed through change-of-state verbs (e.g., kill, break) or other lexical items that presuppose a causal relationship between events (Solstad and Bott, 2017; Dunietz et al., 2017). Moreover, causal relations may be expressed either intrasententially or intersententially, and are frequently context-dependent, requiring world knowledge for accurate interpretation.

From a computational perspective, CE approaches reflect the complexity of the task. Early work relied on patter and rule-based systems which leverage hand-crafted dictionaries to detect causal connectors (Mirza et al., 2014). Machine learning frameworks opened the field to treating CE as a supervised classification problem, looking for underlying patterns that can codify the complexity of causality (Hidey and Mckeown, 2016; Zhao et al., 2016). With the advent of deep learning, Neural Networks were employed to capture local patterns as well as long-distance sequential relationships in causal expression (Kruengkrai et al., 2017; de Silva et al., 2017; Fu et al., 2011; Li et al., 2021; Dasgupta et al., 2018). Most recently, transformer-based models, particularly those fine-tuned on task-specific or domain-adapted corpora, have demonstrated state-of-the-art performance by leveraging contextualized embeddings and selfattention mechanisms to model complex causal semantics (Khetan et al., 2023; Tan et al., 2023; Romanou et al., 2023).

2.2 Domains in CE research

While CE research spans various domains, much of the work to date has concentrated on three areas in particular: scientific literature, news media, and financial documents (Gu et al., 2016; Yu et al., 2019; Mirza, 2021; Tan et al., 2022; Mariko et al., 2021). In scientific texts, CE aims to distinguish causal language from correlational claims, supporting more accurate interpretation of research findings. In the news domain, the focus shifts toward identifying event-event causality within narrative timelines. Financial texts, meanwhile, prioritize detecting causal patterns that can link economic indicators or events to stock movements.

Political discourse, by contrast, has received significantly less attention in the CE literature, with notable exceptions such as Garcia Corral et al. (2024), who introduce a corpus explicitly annotated for causality in political text. Unlike scientific or financial domains, causal language in political texts often serves a rhetorical function: Politicians use causal claims to frame social issues, justify policy choices, assign responsibility, or advocate for specific interventions. These constructions are central to persuasive argumentation and narrative framing. As such, accurate representations of the causal claims made in political discourse are crucial for analyzing policy argumentation (such as in political debates), generating political discourse summaries (for example, in press conferences), and fact-checking claims (important during electoral debates) (Vössing, 2023; Falk and Lapesa, 2022; Reisert et al., 2018).

Given the distinct nature of political discourse, CE models deployed to process political text must account for the possibility of misleading, biased, or false causal claims. Unlike in scientific texts, where extracted causal relations need to correspond to empirical findings, causal statements in political communication may be ideologically motivated and factually inaccurate. For tasks such as automated fact-checking, models must not only detect causal constructions but also preserve their linguistic form in order to accurately represent their argumentative role within disinformation or propaganda for downstream applications.

2.3 CE capabilities of LLMs

As with most other NLP tasks, CE has been probed on LLMs. To evaluate the CE capabilities of LLMs, the task must shift from extraction to generation, aligning with the behavior of these models. Specifically, given a prompt, we expect an output that identifies the cause and the effect of a single sentence.

CE is non-trivial in understanding causal reasoning capabilities in LLMs. Firstly, because we need to test if the models are capable of distinguishing between linguistic causal relationships and real-world causal relationships. When asking a model if the sentence “I couldn’t land in Tokyo because Godzilla took over the runway.” is causal, we expect it to be able to detect that the events “couldn’t land” and “Godzilla took over the runway” are *linguistically* causally connected in the sentence, even if this is not a causal in the real world.

Model: Gemma-7b-It
Prompt: Tell me if this sentence is causal:
I couldn’t land in Tokyo because Godzilla took over the runway.
Output: Sure here is the answer:
The sentence “I couldn’t land in Tokyo because Godzilla took over the runway” is not causal.
It is a humorous fantasy sentence. Godzilla is a fictional creature and does not exist in the real world.

The memorization vs. generalization question presents a complex problem, specially in the political domain which often contains inaccurate claims which may diverge sharply from the factual distributions represented in the model’s training data. For instance, when discussing policies around climate change, political actors may make inaccurate claims about the causes of the climate change. When tasked with extracting causal relations, an LLM may fail to accurately represent the statements, instead defaulting to the causal mechanisms it has learned as “true” from the scientific texts it was trained on. This misalignment can result in outputs that reflect presumed real-world causal relationships rather than faithfully capturing the linguistic causal relationship structure and intent of the original text. In politically sensitive contexts, this raises concerns about the model’s ability to represent, rather than evaluate or correct, linguistic causal relationships as expressed in the source text.

The second challenge involves LLM guardrails.

To mitigate risks of misuse, model creators implement safety mechanisms to restrict model behavior using training-time interventions and post hoc flagging and filtering of inputs and outputs (Wei et al., 2023). For CE this means that when running downstream tasks used to analyze public opinion, models could run into text that is censored by the model because of toxic or unsafe content, potentially hindering a response about the cause and effect span of a flagged sentence. The question raised here is, when using a prompt specifically tailored to undertake a linguistic task, can LLMs interpret the task *literally* and produce an output, or will the safety mechanisms interfere with the task?

2.3.1 State of the art

Recent studies have analyzed the performance of LLMs for CE. In a comprehensive evaluation of ChatGPT's capabilities, Takayanagi et al. (2024) assessed its performance across both domain-specific and non-English datasets. They found that while ChatGPT demonstrates a baseline proficiency in CE, it can be outperformed by earlier models when sufficient training data is available. Moreover, Hobbhahn et al. (2022) explored GPT-3's capacity to identify causes and effects. Their results emphasize the significance of prompting, which suggests that GPT-3's predictions may be influenced more by the form of the input than by its content, raising questions about the model's true understanding of causality. Similarly, Gao et al. (2023) conducted an analysis of ChatGPT's abilities as a causal reasoner. Their experiments suggest that although ChatGPT can provide causal explanations, it struggles with causal reasoning itself, frequently producing "hallucinated" causal connections that do not align with human understanding.

Kıcıman et al. (2023) also tested GPT models and showed that they outperform existing algorithms on tasks such as pairwise causal discovery, counterfactual reasoning, and identifying actual causality. The WIKIWHY benchmark proposed by Ho et al. (2022), aims to differentiate between mere memorization of cause-effect pairs and a genuine understanding of the underlying causal mechanisms. In baseline assessments with GPT-3, just 38.7% of the model's responses were rated as correct by human evaluators. Additionally, Jin et al. (2024) conducted a post-hoc analysis using natural language prompts to describe various causal stories behind X, Y pairs. Their experiments revealed that prompts aligned with the ground-truth data-generating direction achieved the highest zero-shot performance, exceeding that of anticausal prompts by a margin of 2%.

In this paper, we address some of the key challenges identified in the literature: 1) We investigate causal hallucinations across different size and architecture models, and investigate if there is a correlation between model size and causal hallucinations. 2) We examine a new domain that has not been tested for LLMs, political text, and evaluate how LLMs perform in highly rhetorical, low-domain corpora. 3) Using adversarial prompting, we test if models memorize causal pairs rather than truly understand the linguistic relationships, by using sensitive, false and inaccurate claims as evaluation data. 4) We also address temporal factors by studying if models show a decline in performance when parsing events that occurred after their training data had been collected.

3 Evaluation settings

To assess the current limitation in the literature, we broadly test a range of LLM architectures and sizes for CE. As an initial benchmark across the variety of models, we perform a zero-shot test on the **PolitiCAUSE** corpus (Garcia Corral et al., 2024), which comprises sentences from United Nations General Debates annotated for causality. The dataset includes discourse from countries around the world, covering a broad range of ideological positions and themes. We chose this corpus as we expect LLMs to have had prior exposure to similar political texts and themes, which would confer an advantage if the models are solely relying on learned patters from their training data to perform CE.

Furthermore, we designed the task of **Linguistic Causality Disambiguation** (LCD), where we perform adversarial testing by using political fake news that contain sensitive information, and data from out-of-training political events. Based on the task definition of CE provided by Tan et al. (2022) (Sequence classification, span detection, pair classification), our study focuses on the first two: Causal Sequence Classification (CSC) aimed at identifying whether or not a sentence contains a linguistic causal relationship, and Causal Span Detection (CSD) aimed at identifying and differentiating the cause and effect events within the causal sequence.

We address causality extraction at the sentence level. When presented with a sentence S that contains two events through entities e_1 and e_2, the objective is to determine whether there exists a linguistic causal relationship between the entity pair $e_1 - e_2$ within sentence S. Moreover, we account for directionality, as causes can only lead to effects. We label each pair as $e_1 c - e_2 e$, simplifying the label to $cause$ and $effect$.

3.1 Datasets and Evaluation Metrics

3.1.1 Causal Sequence Classification

To evaluate LLMs on CSC in political text, we use a subset of the PolitiCAUSE corpus, filtered to include only sentences containing a complete causal structure (must contain both a cause and an effect span). The evaluation set comprises 527 annotated samples with nearly balanced class distribution (264 non-causal, 263 causal). Each instance includes the original sentence, its binary label, and annotated cause and effect spans. We evaluate model performance using standard classification metrics: Accuracy, Precision, Recall, and F1-score. We are especially interested in Precisions and Recall scores in CSC as high precision and low recall could indicate a reliance of causal markers to identify positive cases while not capturing the full extent of the positive class.

3.1.2 Causal Span Detection

We conducted experiments for CSD using the same data subset as in CSC. To evaluate span detection, we used the SeqEval library (Nakayama, 2018), which calculates the percentage of predictions that exactly match the human-annotated cause and effect spans. We use Precision, Recall and F1-scores, calculated by assessing overlaps between predictions and human labels for each cause and effect. These metrics are calculated on a per word basis. Although SeqEval is a common evaluation framework, interpretation must be done with caution, as span limits are vague in sentences that contain causal claims.

3.1.3 Linguistic Causality Disambiguation

We developed the LCD framework to adversarially test the memorization v.s. generalization questions around CE capabilities of LLMs. We constructed 2 sets of 50 sentences each: 1) **Fake news**, a set that contains real world news that have been flagged as fake news by expert organizations and 2) **Post-training events**, a set containing sentences referring to events that happened after the available knowledge cut-off date of the LLMs (Oct-Dec 2024, Appendix C). With this set, we can analyze the role that the training data has on CE tasks, as well as understand where potential sources of classification errors are coming from. To evaluate the LCD experiments, we use the same evaluation metrics from Section 3.1.1 and 3.1.2. The complete sentence sets are available in Appendix E.

3.2 Experimental Setup

We analyze the performance of LLMs for CSC, CSD and LCD using a zero-shot in-context learning approach. For the first two, we compare a fine-tuned BERT model with zero-shot LLMs to assess whether instruction-following and general pretrained knowledge can substitute for task-specific fine-tuning. A key objective is to examine how robustly each model identifies linguistic causal relationships across political text. Given that modern LLMs are trained with instruction-tuning and alignment techniques, this comparison offers insight into their ability to follow task descriptions without additional supervision. LCD evaluation is only tested using LLMs, to focus on how LLM architecture particularities interfere with CE.

The prompt was created via an iterative process searching for best expected output and looking to maximize the linguistic analysis capabilities on the training set that was not used for the LLMs. The initial version of the prompt was taken from the literature (Takayanagi et al., 2024). Feedback from initial tests using out of sample data led to refinements in wording, structure, and the inclusion of specific linguistic and annotation vocabulary designed to enhance clarity and contextual understanding for the models (O'Connor and Andreas, 2021). Furthermore, the same core prompt was maintained for consistency in the testing framework, only adapting the special tokens, like the end-of-string token. The final prompt can be found in the Appendix A.

For model selection, we use three families of high-performing LLMs for our experiments, looking to maximize parameter size variation when available as to 1) examine when causal mining capabilities appear in LLMs according to size, and 2) further explore the generalization v.s. memorization problem. If the biggest models achieve higher scores in the PolitiCAUSE subset but substantially lowers scores in the fake news or recent events sets, this could be evidence that models are not performing a purely *linguistic* information extraction task.

	Model
OpenAI[1]	GPT-3.5-0125 GPT-4-2024-04-09 GPT-4o-2024-08-06
Meta[2]	Llama-3.1-8B-Instruct, LLama-3.1-70B-Instruct, LLama-3.1-405B-Instruct
Google	Gemma2-9b-it[3] Gemma2-27b-it Gemini-1.5-pro-002[4]

Table 1: Selected model and families, model inference specifications can be found in the Appendix B

	Acc	Prec	Recall	F1
GPT-3.5	0.62	0.76	0.62	0.56
GPT-4	0.83	0.83	0.83	0.83
GPT-4o	0.83	0.84	0.83	0.83
Llama-8b	0.62	0.75	0.62	0.57
Llama-70b	0.74	0.82	0.74	0.73
Llama-405b	0.79	0.82	0.79	0.79
Gemma-9b	0.73	0.51	0.49	0.48
Gemma-27b	0.60	0.52	0.40	0.35
Gemini-1.5	0.76	0.77	0.76	0.76

Table 2: Causal Sequence Classification results for PolitiCAUSE subset. The table includes the values of Accuracy, Precision, Recall and F1 score metrics according to model.

4 Experimental Results

4.1 Causal Sequence Classification

GPT-4o and GPT-4[5] achieved the highest macro F1-score (83%), followed by Llama-405b (79%), while Gemma-27b performed the worst (35%), despite being larger than smaller, better-performing models like Llama-8b (57%) and Gemma-9 (48%). Within the Llama-3.1 family, performance improves with model size, showing clear scaling benefits. GPT models also exhibit strong performance gains from GPT-3.5 to GPT-4/4o (+27%), although their exact parameter sizes are undisclosed. In contrast, Google's Gemma models do not show consistent scaling benefits, likely due to hallucination issues in Gemma-27b, though there is a 28% F1 gain from Gemma-9 to Gemini-1.5. Smaller models tend to favor precision over recall, indicating conservative predictions, while larger models maintain high (>76%) and balanced precision/recall (±1–3%), making them more reliable for causal sequence classification.

4.2 Causal Span Detection

We observe similar performance for CSD across models. GPT-4 and GPT-4o achieve the highest F1-scores (64% and 63%), followed by Gemini-1.5 (58%) and Llama-405b (57%). Notably, Llama-70b (53%) performs comparably to larger models. Smaller models—including Llama-8b, Gemma-9b, and Gemma-27b—score below 40%, with GPT-3.5 trailing at 37%. Across all models, recall consistently exceeds precision, suggesting overprediction likely due to ambiguous span boundaries. Detailed results are provided in Table 6.

4.3 Linguistic Causality Disambiguation

The difference between the binary classification and span detection results of the PolitiCAUSE dataset and the fake news and post-training events sentence sets, allows us to analyze performance differences when the model has to deal with sensitive topics, fake news, conspiracy theory, or scientific inaccuracies, as well as with events that happened after the knowledge cut-off dates.

Fake News For binary classification (Table 3), the models with the largest performance difference compared to the PolitiCAUSE baseline are Llama-405b, with a decrease of 40% in F1-score, followed by Llama-8b, with a 29% difference. From the Open AI models, GPT-3.5 (-19%) suffered the greatest decline, GPT-4 and GPT-4o did not experience a significant diminished performance (9% and 5% respectively), suggesting that the more recent GPT models are better at distinguishing between linguistic causality and real world causality. Gemini-1.5 had the smallest difference in F1-score (-2% points). Interestingly, Gemma-27b improved by 18%, potentially due to it's causal hallucination propensity to be overtaken by safety guardrails. In contrast, span detection average F1-score results were within a 1% point difference when compared to the PolitiCAUSE subset (Table 7).

Qualitative Error Analysis To understand the difference between the binary classification results

---

[5] From this points forward GPT and Gemini models will not include model version (i.e. GPT-3.5-0215 is shortened to GPT-3.5), and Llama and Gemma models will be referenced according to their parameters (i.e. LLama-3.1-8B-Instruct will be Llama-8b, Gemma2-9b-it will be Gemma-9b)

	Fake News Set				Post-training events Set			
Model	Acc	Prec	Recall	F1	Acc	Prec	Recall	F1
GPT-3.5	0.46 (-0.16)	0.72 (-0.04)	0.53 (-0.09)	0.37 (-0.19)	0.62 (0.00)	0.79 (+0.03)	0.60 (-0.02)	0.54 (-0.02)
GPT-4	0.76 (-0.07)	0.77 (-0.06)	0.73 (-0.10)	0.74 (-0.09)	0.74 (-0.09)	0.77 (-0.06)	0.75 (-0.08)	0.73 (-0.10)
GPT-4o	0.78 (-0.05)	0.79 (-0.05)	0.80 (-0.03)	0.78 (-0.05)	0.78 (-0.05)	0.78 (-0.06)	0.78 (-0.05)	0.78 (-0.05)
Llama-8b	0.48 (-0.14)	0.42 (-0.33)	0.36 (-0.26)	0.28 (-0.29)	0.62 (0.00)	0.66 (-0.09)	0.61 (-0.01)	0.58 (+0.01)
Llama-70b	0.62 (-0.12)	0.72 (-0.10)	0.67 (-0.07)	0.61 (-0.12)	0.72 (-0.02)	0.76 (-0.06)	0.71 (-0.03)	0.70 (-0.03)
Llama-405b	0.06 (-0.19)	0.51 (-0.31)	0.44 (-0.35)	0.39 (-0.40)	0.76 (-0.03)	0.76 (-0.06)	0.76 (-0.03)	0.76 (-0.03)
Gemma-9b	0.58 (-0.15)	0.45 (-0.06)	0.42 (-0.07)	0.38 (-0.10)	0.80 (+0.07)	0.82 (+0.31)	0.79 (+0.30)	0.79 (+0.31)
Gemma-27b	0.54 (-0.06)	0.59 (+0.07)	0.58 (+0.18)	0.53 (+0.18)	0.62 (+0.02)	0.79 (+0.27)	0.60 (+0.20)	0.54 (+0.19)
Gemini-1.5	0.74 (-0.02)	0.75 (-0.02)	0.76 (0.00)	0.74 (-0.02)	0.59 (-0.17)	0.59 (-0.18)	0.60 (-0.16)	0.59 (-0.17)

Table 3: LCD CSC results for both Fake News and Post-training events sentence sets. The table includes the values of Accuracy, Precision, Recall and F1 score metrics according to model.

and the span detection results, we conducted a qualitative error analysis. We observe that some outputs were not produced, and null results were provided. The main reasons we find for not producing responses are sentences that have claims that are not backed by scientific evidence (3), sentences that mention conspiracy theories (2), sensitive topics (2) and humorous text (2). Finally, we do observe outputs where sentences are parsed correctly, but include a warning that the claim is false or not backed up by scientific evidence (3), or that it's a conspiracy theory, or a sensitive or complex topic (8).

Furthermore, Gemma-9b and Gemma-27B have the most flagged issues related to fake news (11 cases). We see other interesting behavior from the different models: Gemini-1.5, GPT-4 and GPT-4o were the only models to directly output json output without any extra information, as requested by the prompt. In these, we find 2 cases in the GPT-4 and GPT-4o models where the response was a null output for a positive sequence. We can not conclusively say that this was a guardrail action. However, it occurred with the same 2 sentences, both which do have an explicit causal connector and contain nonfactual information. Moreover, Gemma-9b, Gemma-27b and LLama-405b did not fulfill the request for sentences that mentioned genocide. Gemma-27b did not fulfill the request for a text that mentions biological weapons. Full details of the error analysis are in Table 8.

Post-training events For binary classification (Table 3), the model with the largest performance difference, based on F1-score, is Gemma-9b with an increase of 31% points. The biggest decrease was for the Llama-27b, with a 17% difference between F1-scores. From the Open AI models, GPT-4 suffered the greatest decline (-10%), while GPT-3.5 and GPT-4o did not experience significant diminished performances (2% and 5% respectively). Llama-8b had the smallest different in F1-score, with only a decrease of 1% point.

Span detection results (Table 7) showed substantial improvements compared to the PolitiCAUSE corpus for some models, while others behaved similarly to the fake news results. We observe that all models, expect for Gemini-1.5, show an increase of performance between 6% to 12% in their F1-scores. While Geimini-1.5 had a decrease of 20% F-1 score.

Qualitative Error Analysis The qualitative error analysis did not detect significant issues, given that all models parsed results according to prompt instruction. According to their knowledge cut-off dates (with the exception of the Gemma-2 series, which has no official dates published but online sources suggest June 2024), none of the models should have considered the sentences as real factual information given that the events are not included in their training data. There were only three cases of suspected issues, which involved diseases, the ICC, and the new Mexican president being Jewish, all cases of positive sequence where the models produced 0, null, null (Gemini-1.5 (3), Llama-8 (2), Llama-405(1), and GPT-4 and 4o (3). However due to lack of explanation output from the models, we can not conclusively say that this was due to them being recent events, and the content of the text suggests it could instead be guardrail interference due to sensitive content.

5 Analysis

Our findings consistently demonstrate that larger LLMs exhibit better performance in both CSC and CSD tasks for political domain data. We observed a clear positive correlation between model size and overall performance, indicating that larger models are better equipped to distinguish linguistic causal relationship patterns. Notably, following precision and recall values, evidence suggests that smaller models maybe relying on more explicit causal markers, while larger models demonstrate an ability to identify weaker causal signals. Our experiments did not show widespread causal hallucination issues, although Gemma-27b model did over produce positive cases.

Our span detection results mirror those from the classification tasks, indicating that larger models are more effective at identifying linguistic causal relationships. However, precision and recall metrics reveal that span detection remains challenging due to the inherent complexity of the spans, which are often lengthy and syntactically diverse. The observed recall-over-precision trend suggests that while models are generally adept at recognizing the presence of linguistic causal content, they struggle with accurately delineating span boundaries. This difficulty is likely exacerbated by the ambiguity of the causal claims in naturalistic texts and the tendency of models to favor inclusivity, especially when causal cues are diffuse or embedded within complex syntactic structures.

Furthermore, our experiments on the LCD task reveal that LLMs can exhibit notable interference effects when processing causal claims in politically charged or socially sensitive contexts. While most models are capable of identifying explicit causal claims, the performance disparity between model sizes highlights a key distinction: larger models demonstrate a greater capacity to disambiguate linguistically encoded causal relationships from world knowledge-driven associations. This suggests that scale contributes not only to broader generalization, but also to more faithful alignment with syntactic and semantic cues in discourse. In contrast, smaller models appear more susceptible to heuristic pattern-matching and are prone to overgeneralizing causal signals from pretraining data, particularly in domains such as public health or news, which can contain misinformation.

Moreover, our analysis finds no systematic evidence that model knowledge cut-off dates interfere with CE. Minor performance improvements observed in certain span-level extractions may instead be attributed to variation in textual characteristics across datasets. Specifically, while the PolitiCAUSE dataset comprises utterances from naturally occurring speech (e.g., UN debates), the Fake News and Post Training Events data comes from online news sources with simpler sentence structures and more explicit linguistic causal framing. These stylistic and syntactic differences likely reduce ambiguity in span boundary and contribute to improved model performance.

6 Conclusions

- There is a general correlation between model size and overall performance. Small models can classifying sentences with strong causal signals, while bigger models classify weaker causal signals, potentially relying on their context to process the linguistic signals.

- Results from the LCD experiments show that models for the most part recognize linguistic causal relationships and either produce null results with a warning about the sensitivity of the topic, or parse correctly and include a warning about the content.

- However, we observed a big difference in performance between bigger and smaller models, suggesting that bigger models can better differentiate between linguistic causal relationships and real-world causal relationships, following the task description in the prompt more literally.

- We see no real evidence that models overly rely on training data for the task, as we do not observe that models processing data from the Post-training event set had diminished results.

Taken together, our findings contribute to the growing body of evidence on the linguistic capabilities of LLMs. Specifically, our experiments provide new insights into the performance of LLMs on CE within political discourse, a genre that remains underexplored in NLP and has received limited attention in LLM evaluation. While the results highlight the potential of LLMs in this domain, they also underscore the need for caution when applying these models to politically sensitive texts and relying on their outputs for downstream applications. We argue that more rigorous and targeted evaluation frameworks are essential for understanding how LLMs identify and represent causal claims. By testing models on low-resource, domain-specific text such as political discourse, our work reveals novel aspects of LLM behavior, particularly in their ability to interpret linguistic patterns and follow task-specific prompts.

7 Limitations

Our experiments only consider single-sentence causal sequences, with a complete cause-effect structure, overlooking the complexity of linguistic causality in multi-sentence or cross-sentence contexts, which are prevalent in real-world applications and needed to construct full causal chains of events. Expanding to multi-sentence linguistic causality could provide a more robust evaluation of the models' abilities. The generalizability of the results is limited by the test data size, monolingual setup and the training strategy. Our data is part of real world debates and development that are most likely in the training data, the capabilities of the model beyond these events is not known. Further research is needed to establish if our results generalize beyond English. By addressing these limitations in future work, we can provide a more comprehensive evaluation of LLMs in CE and their broader applications.

8 Ethics Statement

Causal claims can reflect underlying mental models that may be racist, sexist, xenophobic, or derogatory towards specific groups. Additionally the data used here includes a variety of political ideologies and national perspectives, but dominant viewpoints are more likely to take precedence over alternative perspectives. It's crucial to differentiate between identifying these expressions, and endorsing them. The capability to automatically identify the linguistic causal relationships established in political communication allows for more efficient detection of false information, hate speech, and harmful content, which is our objective.

9 Acknowledgments

The authors thank the DFG (EXC number 2055 – Project number 390715649, SCRIPTS) for providing funding; This project has also received funding from the European Union's Horizon Europe research and innovation program under Grant Agreement No 101057131, Climate Action To Advance HeaLthY Societies in Europe (CATALYSE).

References

Tirthankar Dasgupta, Rupsa Saha, Lipika Dey, and Abir Naskar. 2018. Automatic Extraction of Causal Relations from Text using Linguistically Informed Deep Neural Networks. In *Proceedings of the SIGDIAL 2018 Conference*, pages 12–14, Melbourne, Australia. Association for Computational Linguistics.

Tharini N. de Silva, Xiao Zhibo, Zhao Rui, and Mao Kezhi. 2017. Causal Relation Identification Using Convolutional Neural Networks and Knowledge Based Features. *International Journal of Computer and Systems Engineering*, 11(6).

Brett Drury, Hugo Gonçalo Oliveira, and Alneu de Andrade Lopes. 2022. A survey of the extraction and applications of causal relations. *Natural Langauge Engineering*.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104, Valencia, Spain. Association for Computational Linguistics.

Neele Falk and Gabriella Lapesa. 2022. Scaling up Discourse Quality Annotation for Political Science. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*.

Jianfeng Fu, Zongtian Liu, Wei Liu, and Wen Zhou. 2011. Event causal relation extraction based on cascaded conditional random fields. *Pattern Recognition and Artificial Intelligence*, 24(4):567–573.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. Is ChatGPT a good causal reasoner? a comprehensive evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11111–11126, Singapore. Association for Computational Linguistics.

Paulina Garcia Corral, Hanna Bechara, Ran Zhang, and Slava Jankin. 2024. PolitiCause: An annotation scheme and corpus for causality in political texts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12836–12845, Torino, Italia. ELRA and ICCL.

Jinghang Gu, Longhua Qian, and Guodong Zhou. 2016. Chemical-induced disease relation extraction with various linguistic features. *Database*, page 42.

Christopher Hidey and Kathleen Mckeown. 2016. Identifying Causal Relations Using Parallel Wikipedia Articles. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 3:1424–1433. Publisher: Association for Computational Linguistics (ACL).

Matthew Ho, Aditya Sharma, Justin Chang, Michael Saxon, Sharon Levy, Yujie Lu, and William Yang

Wang. 2022. Wikiwhy: Answering and explaining cause-and-effect questions. *Preprint*, arXiv:2210.12152.

Marius Hobbhahn, Tom Lieberum, and David Seiler. 2022. Investigating causal understanding in LLMs. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. 2024. Can large language models infer causation from correlation? *Preprint*, arXiv:2306.05836.

Vivek Khetan, Somin Wadhwa, Byron Wallace, and Silvio Amir. 2023. SemEval-2023 task 8: Causal medical claim identification and related PIO frame extraction from social media posts. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2266–2274, Toronto, Canada. Association for Computational Linguistics.

Canasai Kruengkrai, Kentaro Torisawa, Chikara Hashimoto, Julien Kloetzer, Jong-Hoon Oh, and Masahiro Tanaka. 2017. Improving Event Causality Recognition with Multiple Background Knowledge Sources Using Multi-Column Convolutional Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. *arXiv preprint*. ArXiv:2305.00050 [cs, stat] version: 2.

Zhaoning Li, Xiaotian Zou, and Jiangtao Ren. 2021. Causality extraction based on self-attentive BiLSTM-CRF with transferred embeddings. *Neurocomputing*, 423:207 – 219.

Dominique Mariko, Hanna Abi Akl, Estelle Labidurie, Stephane Durfort, Hugues de Mazancourt, and Mahmoud El-Haj. 2021. The financial document causality detection shared task (FinCausal 2021). In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 58–60, Lancaster, United Kingdom. Association for Computational Linguistics.

Paramita Mirza. 2021. Event causality. In *Computational Analysis of Storylines: Making Sense of Events*, volume 106, pages 106–124. Cambridge University Press.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the TempEval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19, Gothenburg, Sweden. Association for Computational Linguistics.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Joe O'Connor and Jacob Andreas. 2021. What context features can transformer language models use? *CoRR*, abs/2106.08367.

Paul Reisert, Naoya Inoue, Tatsuki Kuribayashi, and Kentaro Inui. 2018. Feasible Annotation Scheme for Capturing Policy Argument Reasoning using Argument Templates. In *Proceedings of the 5th Workshop on Argument Mining*, pages 79–89, Brussels, Belgium. Association for Computational Linguistics.

Angelika Romanou, Syrielle Montariol, Debjit Paul, Leo Laugier, Karl Aberer, and Antoine Bosselut. 2023. CRAB: Assessing the strength of causal relationships between real-world events. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15198–15216, Singapore. Association for Computational Linguistics.

Torgrim Solstad and Oliver Bott. 2017. Causality and Causal Reasoning in Natural Language. In *The Oxford Handbook of Causal Reasoning*. Oxford University Press.

Takehiro Takayanagi, Masahiro Suzuki, Ryotaro Kobayashi, Hiroki Sakaji, and Kiyoshi Izumi. 2024. Is chatgpt the future of causal text mining? a comprehensive evaluation and analysis. *Preprint*, arXiv:2402.14484.

Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza, and Tiancheng Hu. 2022. The Causal News Corpus: Annotating Causal Relations in Event Sentences from News. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*.

Fiona Anting Tan, Xinyu Zuo, and See-Kiong Ng. 2023. UniCausal: Unified Benchmark and Repository for Causal Text Mining. *arXiv preprint*. ArXiv:2208.09163 [cs].

Konstantin Vössing. 2023. Argument-stretching: (slightly) invalid political arguments and their effects on public opinion. *European Political Science Review*, pages 1–21. Publisher: Cambridge University Press.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *Preprint*, arXiv:2307.02483.

Bei Yu, Yingya Li, and Jun Wang. 2019. Detecting Causal Language Use in Science Findings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4664–4674, Hong Kong, China.

Sendong Zhao, Ting Liu, Sicheng Zhao, Yiheng Chen, and Jian-Yun Nie. 2016. Event causality extraction based on connectives analysis. *Neurocomputing*, 173:1943–1950.

A Prompt

Prompt with example output

System: You are a causal language model that performs causal sequence classification and causal span detection. You will classify a text as causal or not causal, and if it's causal you will extract the causes and effects. The output should be a json with label 1 or 0, cause, and effect value such as {\n \"label\": ,\n \"cause\": ,\n \"effect\": \n}.

User: But to pay for it, we had to take on debt, precipitated by massive reduction in Government revenue.

Assistant:{\n \"label\": 1,\n \"cause\": \"massive reduction in Government revenue\",\n \"effect\": \"had to take on debt\"\n}

B Models and model configurations

All Open AI models were run using the Open AI Batch API. Llama and Gemma models were accessed via the Transformers library from Hugging Face, and inference was run using the Together AI API. Gemini-1.5 was run using Google's AI Studio API. Available parameter specifications were established equally across the LLMs: temperature was set to 0.0, Top p to 1.0, Frequency, presence or repetition penalty were all set to 0.0. The maximum number of tokens was specified as 200. For Google models, the `HarmBlockThreshold` in the safety settings parameter was set to `None` for the first two experiments, and set to default for the Fake news and post-training events set of sentences in LCD evaluation. See Table 4 and Table 5 for full details.

Parameter	Value
Temperature:	0.0
Top p:	1.0
Top k:	1.0
Frequency penalty:	0.0
Presence penalty:	0.0
Repetition penalty:	0.0
Max body tokens:	200

Table 4: Model parameter specification. Other parameters that are not included in this table are set to their default values.

C Evaluation processing

To ensure reproducibility, each model's output was subjected to identical pre-processing and post-processing steps. The preprocessing involved token normalization and sequence truncation to maintain consistency across the models. Post-processing included error correction and format standardization. The metrics were selected to provide a comprehensive assessment of model performance across various dimensions of language understanding and generation. Detailed results and further discussion can be found in the subsequent sections of this appendix. We used the SpaCy library to process the text into IOB2 format. For CSD evaluation, we employed our own script available on our Github page. Finally, to analyze classification results, we used the "classification_report" function from Scikit-learn.

If the LLM did not provide a result, as in Null or a text that is not a label, it's was considered a incorrect classification. To analyze the models the missing values (NaN, Null, None, Missing, or empty) are changed to a -1 labeled, and given to the classification report function. For example sentence "So what are we missing?" from the PolitiCAUSE subset (id. 2914) was not parsed by Llama and Gemma models, and it's predicted label was set to -1 before processing the classification report.

D Results tables

Model	Context Window	Knowledge cut-off date	Parameters
		Open AI	
gpt-3.5-0125	16,385	Sep 2021	N.S.
gpt-4-2024-04-09	128,000	Dec 2023	N.S.
gpt-4o-2024-08-06	128,000	Dec 2023	N.S.
		Meta	
Meta-Llama-3.1-8B-Instruct-Turbo	128,000	Dec 2023	∼8B
Meta-Llama-3.1-70B-Instruct-Turbo	128,000	Dec 2023	∼70B
Meta-Llama-3.1-405B-Instruct-Turbo	128,000	Dec 2023	∼405B
		Google	
gemma-2-9b-it	8,192	June 2024*	∼9B
gemma-2-27b-it	8,192	June 2024*	∼27B
gemini-1.5-pro-002	128,000	Sep 2024	N.S.

Table 5: Model details, including the full name used to access the models and their versions. Context window is determined by model provider. Knowledge cut-off dates are provided by AI companies (Dates with * are not official and taken from online sources.) Parameter size is rounded to the next billion (N.S. stands for Not Specified).

	Prec	Recall	F1-score	Prec	Recall	F1-score	Prec	Recall	F1-score
Model		CAUSE			EFFECT			AVERAGE	
GPT-3.5	0.37	0.43	0.37	0.36	0.47	0.38	0.36	0.45	0.37
GPT-4	0.63	0.71	0.64	0.62	0.75	0.64	0.62	0.73	0.64
GPT-4o	0.62	0.70	0.64	0.60	0.75	0.63	0.61	0.73	0.63
Llama-8b	0.40	0.46	0.40	0.41	0.45	0.40	0.40	0.45	0.40
Llama-70b	0.53	0.59	0.53	0.51	0.61	0.52	0.52	0.60	0.53
Llama-405b	0.57	0.63	0.58	0.56	0.64	0.57	0.57	0.64	0.57
Gemma-9b	0.36	0.46	0.38	0.36	0.50	0.39	0.36	0.48	0.38
Gemma-27b	0.36	0.46	0.38	0.36	0.50	0.39	0.36	0.48	0.38
Gemini-1.5	0.58	0.65	0.59	0.56	0.66	0.58	0.57	0.65	0.58

Table 6: Causal Span Detection results for PolitiCAUSE subset. The table includes the values of Precision, Recall and F1 score metrics for each span, and the average score, according to model.

Fake News Set

	CAUSE			EFFECT			AVERAGE		
Model	Prec	Recall	F1-score	Prec	Recall	F1-score	Prec	Recall	F1-score
GPT-3.5	0.28 (-0.09)	0.37 (-0.06)	0.30 (-0.07)	0.33 (-0.03)	0.32 (-0.15)	0.32 (-0.06)	0.31 (-0.05)	0.35 (-0.10)	0.31 (-0.06)
gpt-4o	0.66 (+0.03)	0.73 (+0.02)	0.67 (+0.03)	0.66 (+0.04)	0.66 (-0.09)	0.66 (+0.02)	0.66 (+0.04)	0.70 (-0.03)	0.67 (+0.03)
GPT-4o	0.65 (+0.03)	0.72 (+0.02)	0.67 (+0.03)	0.69 (+0.09)	0.68 (-0.07)	0.68 (+0.05)	0.67 (+0.06)	0.70 (-0.03)	0.67 (+0.04)
Llama-8b	0.36 (-0.04)	0.38 (-0.08)	0.36 (-0.04)	0.39 (-0.02)	0.38 (-0.07)	0.37 (-0.03)	0.37 (-0.03)	0.38 (-0.07)	0.37 (-0.03)
Llama-70b	0.49 (-0.04)	0.55 (-0.04)	0.50 (-0.03)	0.50 (-0.01)	0.51 (-0.10)	0.50 (-0.02)	0.49 (-0.03)	0.53 (-0.07)	0.50 (-0.03)
Llama-405b	0.47 (-0.1)	0.55 (-0.08)	0.49 (-0.09)	0.51 (-0.05)	0.49 (-0.15)	0.49 (-0.08)	0.49 (-0.08)	0.52 (-0.12)	0.49 (-0.08)
Gemma-9b	0.39 (+0.03)	0.50 (+0.04)	0.41 (+0.03)	0.40 (+0.04)	0.39 (-0.11)	0.39 (0.00)	0.40 (+0.04)	0.44 (-0.04)	0.40 (+0.02)
Gemma-27b	0.39 (+0.03)	0.50 (+0.04)	0.41 (+0.03)	0.40 (+0.04)	0.39 (-0.11)	0.39 (0.00)	0.40 (+0.04)	0.44 (-0.04)	0.40 (+0.02)
Gemini-1.5	0.64 (+0.06)	0.69 (+0.04)	0.64 (+0.05)	0.64 (+0.08)	0.65 (-0.01)	0.63 (0.05)	0.64 (+0.07)	0.67 (+0.02)	0.64 (+0.06)

Post-training events Set

	CAUSE			EFFECT			AVERAGE		
Model	Prec	Recall	F1-score	Prec	Recall	F1-score	Prec	Recall	F1-score
GPT-3.5	0.53 (+0.16)	0.51 (+0.08)	0.50 (+0.13)	0.50 (+0.14)	0.49 (+0.02)	0.48 (+0.10)	0.51 (+0.15)	0.50 (+0.05)	0.49 (+0.12)
GPT-4	0.70 (+0.07)	0.72 (+0.01)	0.70 (+0.06)	0.71 (+0.09)	0.71 (-0.04)	0.70 (+0.06)	0.71 (+0.09)	0.71 (-0.02)	0.70 (+0.06)
GPT-4o	0.75 (+0.13)	0.72 (+0.02)	0.72 (+0.08)	0.72 (+0.12)	0.73 (-0.02)	0.72 (+0.09)	0.74 (+0.13)	0.73 (0.00)	0.72 (+0.09)
Llama-8b	0.51 (+0.11)	0.41 (-0.05)	0.44 (+0.04)	0.51 (+0.10)	0.47 (+0.02)	0.48 (+0.08)	0.51 (+0.11)	0.44 (-0.01)	0.46 (+0.06)
Llama-70b	0.63 (+0.10)	0.59 (0.00)	0.60 (+0.07)	0.64 (+0.13)	0.61 (0.00)	0.61 (+0.09)	0.63 (+0.11)	0.60 (0.00)	0.60 (+0.07)
Llama-405b	0.72 (+0.15)	0.67 (+0.04)	0.68 (+0.10)	0.70 (+0.14)	0.69 (+0.05)	0.69 (+0.12)	0.71 (+0.14)	0.68 (+0.04)	0.68 (+0.11)
Gemma-9b	0.52 (+0.16)	0.46 (0.00)	0.47 (+0.09)	0.50 (+0.14)	0.49 (-0.01)	0.49 (+0.10)	0.51 (+0.15)	0.48 (0.00)	0.48 (+0.10)
Gemma-27b	0.52 (+0.16)	0.46 (0.00)	0.47 (0.09)	0.50 (+0.14)	0.49 (-0.01)	0.49 (+0.10)	0.51 (+0.15)	0.48 (0.00)	0.48 (+0.10)
Gemini-1.5	0.40 (-0.18)	0.38 (-0.27)	0.38 (-0.21)	0.40 (-0.16)	0.39 (-0.27)	0.39 (-0.19)	0.40 (-0.17)	0.38 (-0.27)	0.38 (-0.20)

Table 7: LCD CSD results for both Fake News and Post-training events sentence sets. The table includes the values of Precision, Recall and F1 score metrics for each span, and the average score, according to model.

E LCD data and results

Fake News Sentences Set

1. CDC study finds overwhelming majority of people getting coronavirus wore masks.
2. Major Doctor's organization publicly embraces dismembering babies until birth.
3. High doses of vitamin C can cure COVID-19.
4. Chickens are not laying eggs, because RNA is being added to commercial chicken feed.
5. Disney is attempting to lower the minimum drinking age on Disney property to increase their revenue at EPCOT.
6. In January, Spetznas rescued 50 imprisoned, emaciated children from an "Adrenochrome farm" near Shostka, Ukraine.
7. Their captors had so frequently drained their brittle bodies of blood and adrenal fluid that their survival depended on being kept in a medically induced coma.
8. The U.S. Capitol police gave the protesters an "okay" to enter the Capitol.
9. The death was a filmed public execution of a black man by a white cop, with the purpose of creating racial tensions and driving a wedge in the growing group of anti deep state sentiment from common people.
10. Tonight... we move into the residential areas... the white hoods.... and we take what's ours.
11. You will test positive for COVID-19 if you've gotten flu shots during the past ten years.
12. Doctor Stella Immanuel proclaims that a cocktail of hydroxychloroquine, zinc and the antibiotic azithromycin has cured COVID-19 patients.
13. Wisconsin's dairy farms were decimated under the Obama/Biden administration but are doing very well.
14. If Joe Biden's elected, he'll end fracking.
15. Clintons, the Obamas & the Biden family are all involved in child trafficking and crimes against children.
16. Biden went to Ukraine and threatened to withhold $1 billion in aid if they did not fire the prosecutor that was investigating his son and the company that his son worked for.
17. For those wondering if it's worth impeaching him this time, it means he loses his 200K+ pension for the rest of his life.
18. The vaccine contains a spike protein called syncytin-1, vital for the formation of the human placenta.
19. The White House is freaking out after Myanmar Military arrests political leaders for Election Fraud in their November 8 elections.
20. The Trump administration made no effort to get U.S. medical experts into China as the novel coronavirus epidemic spread there early this year.
21. Bernie Sanders calls for a tax rate of 52% on incomes of $29,000 or more to pay for his Medicare for All plan.
22. Dr. Lieber charged for lying about his participation in a Chinese recruitment program and his affiliation with a Chinese university.
23. Do the COVID-19 vaccines cause infertility?
24. The Biden administration just dished out 86 million dollars for motel rooms to house illegal immigrants.
25. Congrats society, we're so woke that we have segregation again.
26. The West is now relying on his wife Julia instead of Alexei Navalny.
27. Germany ran its own military-biological program in Ukraine and was involved in research on biological weapons.
28. Ukraine is committing genocide against the Russian-speaking population in eastern Ukraine.
29. The Bucha massacre was staged by Ukrainian forces.
30. Western governments have imposed sanctions on Russia because of their Russophobia, and Russians are experiencing systematic discrimination in the West.
31. COVID came from China and the vaccine also came from China, don't trust China!
32. Compare this with Russia and China using the pandemic excuse to expand their influence and profit even though the Russian vaccine is ineffective and the Chinese vaccine contains pork gelatin.
33. Muslim scientists from the Raza Academy in Mumbai reported that the Chinese coronavirus vaccine contains gelatin from pork and recommended against vaccination with the haram vaccine.
34. Turkmenistan residents report that the Chinese

vaccine causes severe side effects.

35. Taiwanese travelers stranded at Osaka's Kansai International Airport were rescued by Chinese Embassy.
36. Bolsonaro made a pact with a Masonic sect to win the election.
37. Lula paid his own wife to kill his own son.
38. Brazil Supreme Court Justice Edson Fachin was a lawyer for the MST.
39. Russia unleashed more than 500 lions on its streets to ensure that people are staying indoors during this pandemic outbreak.
40. This document form March 5, from the Government of Spain, forced all nursing homes to keep all seniors with covid-19 symptoms locked in their rooms, keeping them from going to the hospital to get treated.
41. There are 500 boats in Argelia about to sail to Spain... 5,000 illegal immigrants will arrive shortly, many of them infected.
42. Bill Gates owns the patent for the coronavirus.
43. Qatar supported extremist organizations with more than $64 billion over the years!
44. Saudi Crown Prince Mohammed bin Salman had been forced out of power.
45. He announced that Mexican President lowered his own salary and those of 35,000 government employees by 60% to increase pensions for citizens.
46. Any beneficiary of Bolsa Família will lose their benefits if they work as election officials.
47. The Pfizer vaccine produces a fever for 8 to 12 days and after recovery masks are no longer needed.
48. We don't need to continue being named the Republic of Chile or to continue using the Pinochet-imposed flag.
49. Argentine-produced ammunition found in the autopsies of 22 Bolivians killed under Interim President Jeanine Áñez.
50. When you go vote next June 6th, use your own pen or marker, their using all sorts of tricks to commit electoral fraud.

Post-training events Sentences Set

1. Moldovans have received anonymous death threats to scare them from voting.
2. The sharp westward shift in Moldova irked Moscow and significantly soured relations with Chișinău.
3. Spain is still reeling from the deadly impact of its worst flooding disaster in decades where at least 158 people are confirmed dead and dozens are missing.
4. A year's worth of rain fell in eight hours in parts of Valencia on Tuesday.
5. While DANAs aren't unusual in the region, the Mediterranean has seen record-breaking warm waters this summer.
6. This week the German coalition collapsed after Scholz decided to fire some of his key ministers.
7. Scholz's government no longer has a majority in parliament as the Traffic-light coalition collapses.
8. However, the opposition could force Scholz out earlier if they can find a majority for an alternative chancellor.
9. Horrific scenes in Quetta after a suicide bomber targeted passengers waiting to board an express train.
10. We expect many injured after the suicide bombing attack at Peshawar-bound Jaffar Express.
11. Following years of stalemate, the breakthroughs in Baku have now begun, here, at COP29.
12. To enable action, Mr. Babayev identified agreement on a fair and ambitious New Collective Quantified Goal (NCQG) on climate finance as the top negotiating priority for COP29.
13. Mr. Babayev stressed that as the first Paris decade comes to a close, COP29 is a moment of truth that will test our commitment to the multilateral climate system.
14. The ICC has issued an arrest warrant for Benjamin Netanyahu for alleged Gaza war crimes.
15. It is the first time that leaders of a democracy and western-aligned state have been charged by the court, in the most momentous decision of its 22-year history.
16. The United States has been clear that the ICC does not have jurisdiction over this matter.
17. Israel has issued an overnight curfew to people in Lebanon seeking to return to southern Lebanon following the truce.
18. People displaced in the conflict immediately travelled back to southern Lebanon resulting in enormous traffic jams throughout the day.
19. Namibia celebrates elections after first female president is elected.
20. Windhoek is reported to be calm on Wednesday, with neither celebrations nor protests and people carrying on with their normal lives.
21. The South Korean president is facing impeachment vote as defense minister offers to resign.
22. In a shock TV speech on Tuesday, President Yoon Suk Yeol decided to impose martial law.
23. When martial law was briefly declared in South Korea we briefly saw armed soldiers entering the National Assembly.
24. New York City police have launched a manhunt for a masked suspect who gunned down the head of a US medical insurance giant.
25. UnitedHealthcare chief executive was fatally shot outside the Hilton Hotel in Midtown Manhattan.
26. The bells of Notre Dame Cathedral rang for the first time as rebuilding has finalized.
27. Watching the Notre Dame's door swing open to a burst of choir song was such an emotional moment for the thousands of people taking in the display from just outside the cathedral.
28. While the eight bells of Norte Dame were not damaged by the fire, they are ringing for the first time since then.
29. Syrian rebels seized the capital Damascus unopposed on Sunday after a lightning advance that sent President Bashar al-Assad fleeing to Russia.
30. Moscow gave asylum to Assad and his family.
31. The sudden overthrow at the hands of a revolt partly limits Iran's ability to spread weapons to its allies and could cost Russia its Mediterranean naval base.
32. Police don't know how the driver in Magdeburg was able to circumvent the barriers that were protecting the market.
33. We informed the public that a driver plowed a vehicle into a Christmas market killing an adult and a small child.

34. It crashed on the opposite shore of the Caspian after an emergency that was caused by a bird strike.
35. One of the Azerbaijani sources disclosed results that showed the plane was struck by a Russian Pantsir-S air defense system.
36. Russian air-defense system downed Azerbaijan plane.
37. The WHO is aware of the unidentified disease and is sending a team to Congo.
38. The press reported that a flu-like disease that has killed dozens of people over two weeks in Congo is being investigated.
39. The European Union will press ahead with hefty tariffs on China-made electric vehicles even after the bloc's largest economy Germany rejected them, exposing a rift over its biggest trade row with Beijing in a decade.
40. The proposed duties on EVs built in China of up to 45
41. Shares in European carmakers Renault and Volkswagen rose on hopes the tariffs will help them compete with Chinese rivals on their home turf.
42. Her inauguration launches a six-year term during which she will navigate Mexico's all-important relationship with its northern neighbor.
43. Sheinbaum's election broke barriers, as she's the first woman to lead Mexico is also Mexico's first president of Jewish ancestry.
44. Dockworkers at ports from Maine to Texas began walking picket lines early Tuesday in a strike over wages and automation that could reignite inflation and cause shortages of goods if it goes on more than a few weeks.
45. The union wants a complete ban on automation, and it isn't clear just how far apart both sides are.
46. Supply chain experts say consumers won't see an immediate impact from the strike because most retailers stocked up on goods, moving ahead shipments of holiday gift items.
47. Nihon Hidankyo has won the Nobel Price for it's work on nuclear disarmament.
48. This year's Nobel laureates used tools from physics to construct methods for machine learning.
49. The Nobel Prize in Chemistry 2024 is about proteins, life's ingenious chemical tools.
50. In her oeuvre, Nobel laureate, Han Kang, confronts historical traumas and invisible sets of rules and exposes the fragility of human life.

The sentences were manually selected from reputable sources for both Fake News detection and for Real world events. Sources for each sentence are included in the CSV files of each set list.

Qualitative Analysis

Fake News Set Analysis	Cases
GPT-4	**2**
True causal, produces 0, null, null results, no explanation (but other models provide a warning).	2
GPT-4o	**1**
True causal, produces 0, null, null results, no explanation (but other models provide a warning).	1
Llama-8b	**4**
Recognizes causal claim and produces 0, null, null, because it's humorous not factual.	1
Recognizes causal claim, and parses, and warns that it's a complex social issue, sensitive topic or false claim.	1
Recognizes causal claim, and parses, and warns that its not backed by scientific evidence.	1
Unfulfilled requests.	1
Llama-70b	**1**
Recognizes causal claim, and parses, and warns that it's a complex social issue, sensitive topic or false claim.	1
Llama-405b	**4**
Recognizes causal claim and produces 0, null, null, because it's not backed by scientific evidence.	1
Recognizes causal claim, and parses, and warns that it's a complex social issue, sensitive topic or false claim.	2
Unfulfilled requests.	1
Gemma-9b	**11**
Recognizes causal claim and produces 0, null, null, because its not backed by scientific evidence.	2
Recognizes causal claim, and parses, and warns that it's a complex social issue, sensitive topic or false claim.	4
Recognizes causal claim, and parses, and warns that it's not backed by scientific evidence.	2
Unfulfilled requests.	1
Warns it's a conspiracy theory produces 0, null, null.	2
Gemma-27b	**11**
Recognizes causal claim and produces 0, null, null because its a sensitive topic.	2
Recognizes causal claim and produces 0, null, null, because it's humorous not factual.	1
Recognizes causal claim and produces 0, null, null, because its not backed by scientific evidence.	3
Warns it's a conspiracy theory produces 0, null, null.	5
Gemini-1.5	**3**
True causal, produces null results, no explanation (but other models provide a warning).	3

Post-training events Set Analysis	Cases
GPT-4	**3**
Produced null results for a sentence is positive about a sensitive topic.	3
GPT-4o	**3**
Produced null results for a sentence is positive about a sensitive topic.	3
Llama-8b	**2**
Produced null results for a sentence is positive about a sensitive topic.	2
Llama-405b	**1**
Produced null results for a sentence is positive about a sensitive topic.	1
Gemini-1.5	**3**

Table 8: Qualitative analysis for LCD

Heroes, Villains, and Victims: Character Narratives in the WPS Agenda of the UNSC

Hannah Steinbach **Imge Yüzüncüoglu** **Raluca Rilla** **Manfred Stede**
University of Potsdam
Department of Linguistics
Potsdam, Germany
`{hannahsteinbach0312, i.yuezuencue, ralucarilla}@gmail.com`
`stede@uni-potsdam.de`

Abstract

We investigate how heroes, victims, and villains are constructed in debates on the United Nations Security Council's Women, Peace, and Security (WPS) agenda. Drawing from 2,566 speeches delivered between 2000 and 2019, we examine how (gendered) entities are framed within diplomatic discourse using topic modeling, clustering, and supervised learning. To assess the potential of automated character role identification, we manually annotate 54 speeches with character role labels and evaluate a fine-tuned RoBERTa classifier alongside two chat-optimized Large Language Models (DeepSeek-R1, Llama3.3 70B). Our findings reveal substantial variation in model performance, with RoBERTa demonstrating best overall performance. Our analysis shows that women are framed as both empowered agents and vulnerable subjects, while perpetrators often remain unnamed. The UNSC often casts itself as a hero by emphasizing its contributions to the WPS agenda. All code and annotated datasets are publicly available to facilitate further research on narrative framing and role attribution in this domain.

1 Introduction

According to Gehring and Grigoletto (2023), "narratives are a crucial group-based mechanism that influences human decision-making" (p. 1). The *Narrative Policy Framework (NPF)* (Jones and McBeth, 2010) defines narratives through distinct components including context, storyline, moral of the story, and characters. Building on the drama triangle (Karpman, 1968), the character roles *hero*, *victim*, and *villain* are central elements of narrative construction. While prior research has applied these frameworks extensively in areas such as climate change communication (e.g., Wolters et al. 2021; Fremann et al. 2023; Gehring and Grigoletto 2023; Grasso et al. 2025), less attention has been given to their role in diplomatic settings.

In multilateral settings like the United Nations (UN), the strategic use of narratives allows states to construct legitimacy and influence international norms. By portraying different entities as heroes, victims, or villains, speakers subtly shape perceptions of agency and responsibility. These narrative strategies play a crucial role in the Women, Peace, and Security (WPS) agenda of the UN Security Council (UNSC). Initiated by Resolution 1325 (2000), the agenda seeks to prevent violence against women and girls in (post-)conflict contexts, promote gender equality, increase women's participation in security processes, and integrate a gender perspective (UNDP, 2023). Given these objectives, one can anticipate that women are central figures in these debates, yet they are also framed in specific roles. They often appear as victims, needing protection, yet also as heroes whose (potential) participation in UN peace and security efforts is framed as beneficial. At the same time, the very call to increase their involvement implies that they are still largely excluded from these processes, reinforcing their victimhood. The UN itself is also cast as a hero, both as protector and enabler of women's involvement. Since the objectives of the UNSC are a key component of the agenda, the sessions themselves can be expected to involve similar narratives.

This paper investigates how heroes, victims, and villains are constructed in WPS open debates, and evaluates the ability of Large Language Models (LLMs) to automatically identify these roles. We begin with a meta-analysis of gender representation among UNSC speakers, followed by topic modeling using clustering techniques and BERTopic (Grootendorst, 2022) to explore dominant themes and the framing of gendered entities. To assess the potential for automated classification of narrative roles, we frame the task as a supervised learning problem and fine-tune a RoBERTa model. We then compare its performance to that of two chat-

Proceedings of the 5th Workshop on Computational Linguistics for the Political and Social Sciences (CPSS), pages 48–67
September 10-11, 2025 ©2025 Association for Computational Linguistics

optimized LLMs: Llama and DeepSeek. The manually annotated data serves as a benchmark. Our contributions include:

1. A meta-analysis of gender representation in UNSC WPS debates.
2. An analysis of the most frequent topics in the WPS agenda using clustering methods, combined with a comparison of how gendered entities are discussed across these topics.
3. A new corpus of 54 speeches annotated for hero, victim, and villain roles, with guidelines.
4. Evaluation of automatic role classification using RoBERTa, Llama, and DeepSeek.

All code and annotated datasets are publicly available and encourage further research on narrative framing and role attribution in this domain: GitHub.

2 Related Work

2.1 Narrative Framing in the United Nations

Gibbings (2011) states that "operating at the UN is akin to acquiring a second language" (p. 533), reflecting the cultural norms that shape diplomatic discourse. One of these norms is the expectation to frame issues positively: "UN speech styles encourage positive visions and utopian dreams" (Gibbings, 2011, p. 534). Based on four months of fieldwork on Resolution 1325, Gibbings notes that women are often framed as peacemakers and knowledge-bearers, while framings that deviate from this narrative are discouraged. She illustrates this with the case of two Iraqi women who, at a 2003 UN meeting, criticized the US and UK invasion and the UN's lack of support. Though they were barred from formally speaking, their critical remarks at the informal session led to them being labeled "angry." This incident reveals not only the discursive etiquette within the UN but also that when women display agency, they are expected to do so as peaceful agents.

While the UN frequently depicts women as (peaceful) "agents of change," this framing is often accompanied by portrayals of women as victims, particularly in contexts emphasizing vulnerability and protection. The UN's tendency to depict women primarily through these dual lenses has been widely criticized. Başer (2024), for instance, warns that these framings reinforce gender norms by failing to acknowledge women also as perpetrators of violence and by erasing male or LGBTQ+ victims. Similarly, Carpenter (2005) finds that the UN's civilian protection discourse heavily relies on the phrase "women and children," sidelining men. This framing emphasizes perceived vulnerability and helps garner donor support (Carpenter, 2005).

De la Rosa and Lázaro (2019) analyze metaphorical framings in UNSC resolutions from 2000 to 2015 and find a shift from portraying women solely as victims to depicting them as empowered agents in the fight against sexual violence. This suggests a change in the portrayal of women, at least within the resolutions.

Open debates, however, differ from resolutions. While resolutions are negotiated and largely agreed upon, debates reflect a broader range of Member State perspectives and are less curated. Although True and Wiener (2019) do not explicitly employ the victim, villain, and hero framework, their analysis offers valuable insights consistent with these characterizations. In their study of the 15th-anniversary open debate on WPS in 2015, they find that portrayals of women remain largely victim-focused, particularly in relation to sexual violence, thereby risking the reinforcement of stereotypes. They also identify states positioning themselves as heroes: either as champions of the agenda (e.g., Global North states) or as implementers under constraint (e.g., conflict-affected states). Others, like Russia and China, resist expansion of the agenda, emphasizing sovereignty. Perpetrator framing remains politically delicate. While non-state actors like ISIS are widely condemned, direct criticism of fellow states is rare, since it is easier to call out "breaches of the norm by non-state actors [...] than to name and shame breaches made by states present" (True and Wiener, 2019, p. 563).

True and Wiener (2019)'s analysis of an open debate offers important insights into character framings, highlighting the victimization of women (and, in some cases, states), self-portrayals of heroism by Member States, and the targeting of non-state actors as perpetrators, while notably avoiding direct criticism of present states. Given the UN's tendency to maintain a positive tone, we anticipate an emphasis on achievements and successes—thus, on heroes. Building on these findings, our next step is to assess whether these insights persist across a broader set of open debates and to explore the potential of automated approaches to process large volumes of text more efficiently. The following section reviews existing approaches to automatic character framing analysis

2.2 Automated Character Role Identification in NLP

The automated identification of character roles such as heroes, victims, and villains has long been relevant in NLP, particularly in framing analysis. Gomez-Zara et al. (2018) developed a browser extension that identifies role framings in news articles based on named entity recognition, sentiment analysis, and lexical similarity to prototypical roles. While the implementation has not undergone a formal evaluation yet, the authors acknowledge that the browser extension may struggle with complex sentence structures (Gomez-Zara et al., 2018).

Sap et al. (2017) examine how stereotypes are formed through language by extending the connotation frame lexicon of Rashkin et al. (2016) with verbs that capture "power" and "agency:" "power" reflects the level of authority conveyed by a verb, while "agency" indicates how much a character is “powerful, decisive, and capable of pushing forward their own storyline” (Sap et al., 2017, p.4). To test this empirically, they analyze gender bias in modern films by comparing how male and female characters differ in attributed power and agency. Their approach offers a useful tool for studying narrative structure and power dynamics.

Stammbach et al. (2022) reformulate character role labeling as a reading comprehension task. Prompting GPT-3 with context and role-specific questions (e.g., “Who is the hero?”), they evaluate performance across diverse domains: news articles on fracking, Disney film summaries, and US State of the Union addresses. GPT-3 outperforms rule-based baselines (Gomez-Zara et al., 2018), and proves to be especially effective on narrative text.

Character framing has been frequently studied in the context of ecology and climate change. The developers of the *Character-Role Narrative Framework* (Gehring and Grigoletto, 2023), which builds on the drama triangle (Karpman, 1968), applied it to study US climate policy on Twitter as an example of large-scale climate change discourse, revealing how narrative roles influence the virality and public reception of environmental messages (Gehring and Grigoletto, 2023). For their analysis, they used RoBERTa (Liu et al., 2019) combined with XGBoost (Chen and Guestrin, 2016), concluding that integrating LLMs with a limited training dataset is a "very efficient approach to measuring narratives in large text data sets" (Gehring and Grigoletto, 2023, p. 32).

Following this approach, Grasso et al. (2025) evaluate four LLMs (GPT-4o, GPT-4, GPT-4-turbo, and LLaMA-3.1-8B) for character role detection (adding the class "beneficiary") and categorization on tweets from the *Ecoverse* dataset and editorial paragraphs from *Nature & Science*. They then adapt the best performing model, GPT-4, to create two new annotated datasets of tweets and editorial paragraphs, followed by a subsequent error analysis. They conclude that LLMs are in fact suitable for large-scale character-role analysis in the environmental domain, though they advise that human validation remains necessary.

While previous research has demonstrated the potential of LLMs for role detection in narrative, journalistic texts, and social media posts such as tweets, their applicability to formal diplomatic discourse remains underexplored.

3 Dataset and Task

3.1 Data

Using the dataset of UNSC debates from 1997–2019 by Schoenfeld et al. (2019), we identified 68 sessions (4613 speeches) addressing the WPS agenda here. We removed presidential speeches (which are mainly formalities), the duplicate of one debate, and the proceedings of S/PV.6180 from 2009 under the meeting of the year 2001, after observing that there was no WPS agenda meeting in 2001. The final dataset consists of a total of 2566 speeches, for which we each extracted metadata including year, session, speaker affiliation, gender (based on honorifics), and language of the speech using regular expressions.

3.2 Manual Annotation of Dataset

We manually annotated 54 speeches for the roles of "hero," "victim," and "villain," taking into account both commonly held understandings of these roles and the definitions provided by Klapp (1954), Cohen (2011), Gomez-Zara et al. (2018), Bergstrand and Jasper (2018) and Jasper et al. (2018). The full character definitions can be found in Section Appendix C. The speeches were sampled to roughly reflect their distribution across different years. Based on four pilot annotations, we developed corpus-specific guidelines. In addition to excluding abstract entities (e.g. "sexual violence"), we expanded the definitions of the roles to adapt them

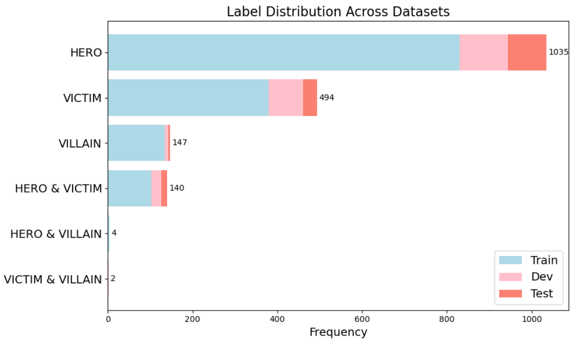


Figure 1: Character Label Distribution.

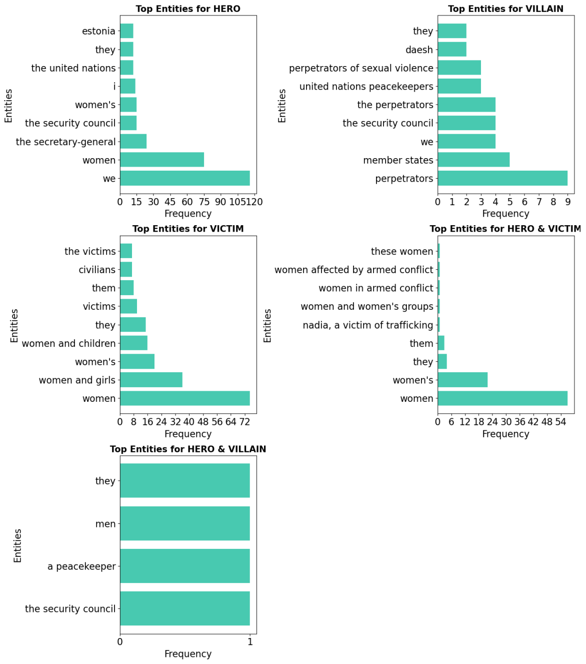


Figure 2: Top Entities per Label

to the context of the data. Two important additions were made: the "hero" category was extended to encompass actors as having potential to bring about positive change but facing structural oppression, and the "villain" category to include actors portrayed as obstructing equal rights and justice for victims. Following Fremann et al. (2023), who criticizes existing datasets and frameworks for assuming that entities can only be assigned a single role rather than recognizing that they can be associated with multiple roles simultaneously, we allowed overlapping roles and nested annotations.

The 54 annotated speeches were split into 44 training set, 5 development set, and 5 test set speeches. Annotations were done individually by the authors using INCEpTION (Klie et al., 2018). Inter-annotator agreement averaged at 0.38 (Krippendorff's α).[1]The relatively low initial agreement stemmed partly from differing interpretations of role definitions, which prompted a subsequent refinement of the guidelines. It was also influenced by smaller inconsistencies in span selection, such as whether to include punctuation, modifiers, or complete noun phrases. These discrepancies reflected the inherent subjectivity and complexity of role labeling task. Nevertheless, after collaboratively reviewing all speeches and discussing disagreements, we reached full consensus on the final labels.

3.3 Annotated Data Insights

The annotation of the 54 speeches offers valuable insights. The following statistics were generated only after model evaluation to avoid introducing any bias.

Figure 1 shows the frequency of roles within the annotated dataset. The "hero" label appears most frequently—twice as often as "victim" and seven times more than "villain," aligning with previous research indicating that speeches at the UNSC are generally framed positively. Figure 2 shows the top entities per label. As expected, "women" dominates the "victim" category, alongside group mentions like "women and girls" and "women and children." For "hero," the most common entity is "we," referring to speakers, Member States, or the Council itself, supporting findings that states emphasize their own contributions (True and Wiener, 2019). Interestingly, "women" also ranks second among heroes, indicating that women are not only portrayed as victims but also credited for their agency and/or potential. This dual role is echoed in the "hero-victim" category, which includes mentions like "women," and "Nadia, a victim of trafficking."

The "villain" label appears much less frequently and is often assigned to vague actors such as "perpetrators" or "perpetrators of sexual violence." In some cases, the Council villainizes itself:

(1) "Such debates remain necessary, considering how far **we** are from full implementation."

As noted earlier, we broadened the "villain" category to include entities that impede equal rights. In this example, the Council acknowledges falling short of its own goals regarding women's equal rights. While this self-criticism is subtle, it reflects a characteristic diplomatic strategy where direct

---

[1] A detailed overview of pairwise agreement scores between individual annotators and our gold labels can be found in Appendix B.

blame is often avoided or softened. This example highlights that the way character roles are defined and assigned can vary across domains, which must be carefully considered when developing automated classification methods.

Aside from this, only two specific villain groups appear: Daesh (ISIS) and UN peacekeepers. The hero-villain dual label is rare but includes references to entities such as "the Security Council" and "a peacekeeper," indicating that while peacekeepers are sent with good intentions, they are sometimes also perpetrators of (sexual) violence. Although it is often implied that perpetrators are male, specific individuals or groups are very rarely identified.

Over time, portrayals of "women" fluctuate (Appendix, Figure 5) but do not show a clear shift from victimhood to heroism as observed in resolutions (de la Rosa and Lázaro, 2019). "Men" are mentioned far less frequently and are often referenced alongside women as perpetrators, allies, or co-victims (Appendix, Figure 6). Mentions of male victims without women or children are rare, reinforcing the framing of "women and children" as the prototypical vulnerable group (Carpenter, 2005). Mentions of male victims increase slightly after 2012, while portrayals of men as perpetrators decline (Appendix, Figure 7). This could suggest increasing recognition of male victims of armed conflict, as well as a reduction in the portrayal of men as perpetrators.

These patterns confirm our core hypotheses: women are primarily framed as victims, though also as heroes; villains remain vague or externalized; and men are largely secondary, often mentioned in relation to women and/or children.

4 Exploratory Work

4.1 Meta-Analysis

Since a core goal of the WPS agenda is to promote the inclusion of women in peace and security processes not only as subjects of protection but as active participants, we examined the representation at the top: within the UN Security Council itself. This led to our first research question: "What does the gender representation of the UNSC look like?"

Using collected meta-information, Figure 3 shows that from 2000 to 2019, women never delivered more than half of all speeches. Female representation increased from under 20% in 2000 to nearly 40% in 2019, peaking just above 40% in 2018. However, this trend is inconsistent and fluctuates across years. Among the permanent members (Appendix, Figure 8), the US stands out with over 80% of speeches delivered by women. The UK follows at just ~20%, followed by France (~10%) and Russia (~2%), while China had no female speakers at all. This imbalance raises questions about the relationship between representation and discourse; though a systematic comparison by country is beyond the scope of this study, it remains a promising direction for future work.

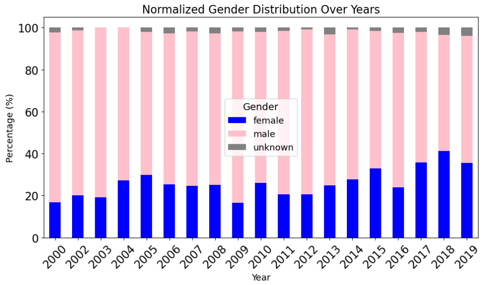


Figure 3: Number of speeches by gender across years. Speakers without gendered honorifics or titles are marked as "unknown."

4.2 Topics

4.2.1 Bertopic

To analyze how gendered entities are framed, we applied BERTopic (Grootendorst, 2022) to all sentences across the corpus that mention either unambiguous female terms (“women”, “girls”) or male terms (“men”, “boys”), excluding sentences where both co-occur. BERTopic generated significantly more topics for female entities (309) than for male ones (7). The number of topics for female mentions were reduced to nine with an in-built function for interpretability and selected topics visualized using word clouds (see all word clouds in Appendix E.3).

Female entities. Some topics focused on institutional discourse, with terms like “director,” “delivered,” and “Ms.” Others centered on the WPS agenda itself, highlighting keywords such as “resolution,” “peace,” and “UN-Women”. Several topics emphasized women as victims, particularly of sexual violence (Figure 9a), human trafficking (Figure 9b), and domestic abuse (Figure 9c).

One topic links women to peacekeepers (Figure 9e), framing them both as heroes via calls for greater female participation and as victims of sexual abuse by (implicitly male) peacekeepers. The following quotes highlight this dual framing:

(2) “We are convinced of the advantages inherent in increasing the number of women Blue Helmets"

(3) “[W]e must take every possible step to root out sexual exploitation and abuse by peacekeepers”

Male entities. Topics related to male entities were fewer and less thematically coherent. One addressed sexual violence (Figure 9f), framing men as both perpetrators and victims:

(4) “[...] the role of men as perpetrators but also as victims of sexual violence in conflict.”

Other topics focused on men’s engagement in gender equality and leadership (Figures 9g–9h), blending heroism, advocacy, and power.

Overall, women are frequently framed as victims but also as (potential) heroes in peacekeeping contexts. Men are mentioned less often and usually in relation to women—as perpetrators, allies, or fellow victims. Villain roles are rarely explicit and often assigned to non-state actors. This implicitness complicates role detection, especially when relying solely on gendered search terms.

4.2.2 Extracting Embeddings and Clustering

We generated 1024-dimensional embeddings for each speech using mxbai-embed-large (Li and Li, 2023), then applied UMAP (McInnes et al., 2018) for dimensionality reduction, and DBSCAN (Ester et al., 1996) for its ability to handle undefined cluster counts and outliers in high-dimensional data. After parameter tuning (eps=6.5, min_samples=5), we identified 18 distinct clusters (see Appendix, Figure 10), revealing two key patterns.

First, we found 14 country-specific clusters, suggesting both consistent national narratives across multiple sessions and standardized diplomatic templates. Beyond typical formalities, we also notice entire phrases such as “*Women are formidable negotiators, mediators, and peacebuilders*” and “*Sexual violence is not cost-free*” are reproduced across multiple speeches, raising questions about the performative versus substantive nature of these statements. Secondly, a smaller subset of clusters grouped together thematically similar speeches. One such cluster centers around sexual violence in armed conflict, with most speakers framing it as a crime against humanity and a threat to peace and security, emphasizing the continued need to take action. Nevertheless, this cluster also includes a speech by the Russian delegation in which the speaker is solely expressing concern over the expansion of UN mandate language beyond existing Security Council resolutions, highlighting that even within thematically coherent clusters, substantive differences in framing exist.

While this approach offers a scalable way to detect patterns in country discourse, it fails to capture *how* countries frame issues like sexual violence, obscuring speaker intentions and level of assertiveness. We therefore turn to a more fine-grained analysis of character-role annotations using (1) a fine-tuned RoBERTa model and (2) direct prompting of LLMs to complement our topic-modeling and clustering approaches.

4.3 RoBERTa Model

Various BERT models were developed by participants in a similar shared task on detecting whether characters in memes were glorified, vilified, or victimized (Sharma et al., 2022). Both the top-performing system (Kun et al., 2022) and the third-ranked system (Singh et al., 2022) incorporated RoBERTa (Liu et al., 2019). Inspired by these works and Gehring and Grigoletto (2023), who also used a RoBERTa model, we fine-tuned RoBERTa for our task.

4.3.1 Preprocessing steps

We adapted the NER task using the standard Beginning-inside-outside (BIO) tagging format (Tjong Kim Sang and Buchholz, 2000) (see Table 5 in the Appendix for an example). Overlapping entities were merged into single combined labels. Since such overlaps were relatively rare (see Figure 1), maintaining them as distinct tags would not have yielded reliable model performance. This resulted in a total of 13 labels[2].

4.3.2 Implementation

To fine-tune the RoBERTa model for the multi-class classification task, we encoded labels and pre-tokenized the input texts using pre-trained RoBERTa embeddings. Model selection was based on performance on the dev set. Details on hyperparameter configurations, performance (for m1-m4), and the rationale for ultimately selecting model m4 for evaluation on the test set are provided in Appendix D.2.

4.3.3 Evaluation & Interpretation

We used the default mode of `seqeval` (Nakayama, 2018), a framework designed for the evaluation of sequence labeling tasks, thus allowing some leniency in evaluation.

---

[2] Labels: HERO, VICTIM, VILLAIN, HERO_VICTIM, HERO_VILLAIN, VICTIM_VILLAIN, each marked with B- and I-prefixes.

	P	R	F1	S
HERO	**0.69**	0.76	0.72	95
VICTIM	0.67	**0.81**	**0.73**	32
VILLAIN	0.43	0.75	0.55	4
HERO_VICTIM	0.36	0.33	0.35	12
micro avg	0.65	0.73	0.69	143
macro avg	0.54	0.66	0.59	143
weighted avg	0.65	0.73	0.69	143

Table 1: Fine-tuned RoBERTa results: precision (P), recall (R), F1, and support (S) by label, plus micro, macro, and weighted averages.

The performance of the RoBERTa-based classifier is summarized in Table 1. The model performs best for the HERO and VICTIM classes, achieving F1 scores of 0.72 and 0.73, respectively. For the VILLAIN class, it shows a high recall (0.75), but lower precision (0.43) and F1 (0.55) scores, indicating frequent false positives. We acknowledge that this is largely due to the fact that there are only four occurrences of label. Most likely, performance would improve with more training data for this category.

The classifier struggles most with the overlapping label HERO_VICTIM (F1: 0.35), while the remaining overlapping classes HERO_VILLAIN and VICTIM_VILLAIN are not represented in the test set. The low performance scores suggest difficulty in distinguishing overlapping roles, likely due to ambiguity in textual representations.

The identical micro and weighted F1 scores (0.69) indicate strong overall performance, especially on frequent classes like HERO and VICTIM. The lower macro-average (0.59), however, highlights class-specific performance gaps, again highlighting that the model struggles with rarer labels. Addressing this imbalance (e.g., through data augmentation, different sampling techniques, or loss function reweighting) could improve performance on underrepresented roles and therefore result in a more balanced model.

Overall, our fine-tuned RoBERTa model performed well, with its predictions closely aligning with our annotations. The consistency of its outputs suggests a high degree of reliability in capturing the intended role assignments. Interestingly, the model also identified some entities that we overlooked in our annotation process, indicating its potential to enhance annotation quality and reduce human error. Therefore, the model could serve as a valuable tool for automating or assisting in large-scale annotation efforts, provided that occasional inconsistencies are accounted for through further refinement or human verification.

4.4 LLMs

Following Stammbach et al. (2022), we evaluated the outputs of two LLMs in zero- and few-shot settings: DeepSeek-R1 (Guo et al., 2025) and Meta-Llama-3-70B (Grattafiori et al., 2024).[3]

DeepSeek, designed for reasoning, is enhanced through pure reinforcement learning, built on the fine-tuned DeepSeek-V3-Base model, and utilized with a reinforcement learning framework. Llama, a state-of-the-art open-weight transformer-based model developed by Meta, has demonstrated strong generalization capabilities, making it a suitable candidate for comparison (Grattafiori et al., 2024).

While we acknowledge that a strict comparison between the two LLMs that we utilized is not entirely fair in virtue of the immense difference in parameters, we still like to comparatively report differences in their performance.

4.4.1 Preprocessing

We prompted models to reproduce the input speech with in-line role tags (e.g., <HER>Ambassador Powers</HER>). We then converted both model outputs and gold annotations into span-based JSON files for comparison. Due to inconsistencies within the gold annotations (e.g., whitespace, newline characters), we implemented a character-level alignment process to ensure accurate offset matching between model predictions and gold spans. Using fuzzy matching (`difflib.SequenceMatcher`[4]) with a similarity threshold of 0.98 and offset mapping adjustments, we recalibrated span boundaries to enable precise comparison.

4.4.2 Implementation

The prompts were iteratively refined with ChatGPT using the dev set, ensuring none of the models chosen for testing has an advantage with prompts optimized to its understanding. Each prompt included the task description followed by the annotation guidelines (Appendix C), a defined template for the output, and a short repetition of the task to support long-text understanding. The few-shot prompt also includes two annotated speeches from the train set. To see the full prompts, please refer to our GitHub repository.

---

[3] For readability, we refer to DeepSeek-R1 and Meta-Llama-3-70B as "DeepSeek" and "Llama," respectively, throughout the paper.

[4] `https://github.com/python/cpython/blob/3.13/Lib/difflib.py`

4.4.3 Evaluation & Interpretation

Table 2 presents the performance of LLMs on the test set using partial match evaluation, to stay in line with our reporting of more lenient evaluation results from RoBERTa. Partial matches required a span overlap of $\geq 50\%$ and a one-to-one mapping between predicted and gold spans, penalizing outputs that over-segment spans.

Label	Model	P	R	F1
HERO	DeepSeekZero	0.39	**0.76**	0.52
	DeepSeekFew	**0.48**	0.72	**0.58**
	LlamaZero	0.29	0.59	0.39
	LlamaFew	0.36	0.65	0.46
VICTIM	DeepSeekZero	0.49	0.65	0.56
	DeepSeekFew	**0.74**	**0.67**	**0.71**
	LlamaZero	0.4	0.65	0.5
	LlamaFew	0.43	0.65	0.52
VILLAIN	DeepSeekZero	0.33	0.25	0.29
	DeepSeekFew	0.67	0.5	0.57
	LlamaZero	**0.75**	**0.75**	**0.75**
	LlamaFew	0.67	0.5	0.57
Micro avg	DeepSeekZero	0.42	0.69	0.52
	DeepSeekFew	**0.54**	**0.7**	**0.61**
	LlamaZero	0.33	0.61	0.43
	LlamaFew	0.38	0.65	0.48
Macro avg	DeepSeekZero	0.4	0.55	0.45
	DeepSeekFew	**0.63**	0.63	**0.62**
	LlamaZero	0.48	**0.66**	0.55
	LlamaFew	0.49	0.6	0.52
Weighted avg	DeepSeekZero	0.44	**0.69**	0.53
	DeepSeekFew	**0.58**	0.66	**0.62**
	LlamaZero	0.36	0.63	0.46
	LlamaFew	0.42	0.61	0.5

Table 2: Classification results for LLMs across role labels. **DeepSeekZero** and **LlamaZero** refer to zero-shot prompting; **DeepSeekFew** and **LlamaFew** use few-shot prompting.

Unlike RoBERTa, which used composite class labels to represent multi-label annotations, we treated each role label independently in the LLM setup. This approach preserved annotation granularity and improved evaluation fidelity by crediting individual label matches.

As shown in Table 2, DeepSeek generally outperforms Llama, particularly in the few-shot setting: For both HERO and VICTIM, DeepSeekFew achieves the highest F1 scores (0.58 and 0.71), with DeepSeekZero yielding the highest recall for HERO (0.76 vs. 0.72 for DeepSeekFew). However, for VILLAIN, LlamaZero performs best (Precision: 0.54, Recall: 0.75, F1: 0.75), though this is just based on four test instances and may not generalize. In future work, this pattern should be investigated with a more balanced dataset.

Overall, DeepSeekFew yields the highest scores across all aggregates: micro-F1 (0.61), macro-F1 (0.62), and weighted-F1 (0.62). These results suggest better generalization and handling of class imbalance compared to Llama. DeepSeekFew achieves a slightly higher macro-average F1 (0.62) than RoBERTa (0.59). This may be due to our decision to treat each role label independently in the LLM setup, avoiding multi-label combinations (RoBERTa's lowest scoring class). However, both LLMs generally underperform relative to RoBERTa. In the next section, we analyze how and why these models fail.

4.5 Error Analysis

Noun Phrases, Articles and Pronouns The tested LLMs often struggled to capture full NPs, often separating phrases such as "women and children." Moreover, articles, pronouns, and descriptive modifiers were frequently omitted (e.g., "thousands of" in "thousands of women and girls"). Additionally, some entities were overextended to include abstract concepts (e.g., "the zero tolerance policy of the secretary general" instead of "the secretary-general"). Importantly, few-shot prompting slightly improved consistency, suggesting that prompt length and structure may impact model stability (Wu et al., 2024).

Mislabeling Formal Entities DeepSeekZero regularly incorrectly labeled speakers as heroes. DeepSeekFew frequently mislabeled thanked individuals as heroes, while RoBERTa only occasionally did this, and Llama never did. Resolutions were also frequently mislabeled as heroes by both DeepSeek models and LlamaFew. LlamaFew further misclassified "the Dakar Declaration" as a hero, although we instructed models against labeling non-resolution documents.

Bias Several models displayed patterns suggesting biases, such as RoBERTa tending to over-label official entities (e.g. "council," "the Secretary-General," "my government") as heroes. More concerning is that RoBERTa often assigned the role of victim to women without justification. LlamaZero demonstrated a similar problem: In the phrase "training must target not only girls and women, but also boys and men"[5], it assigned both "hero" and "victim" to "girls and women," while labeling "boys and men" as only "hero." In the following example, all models annotated women only as victims, rather than as both victims and heroes:

(5) "[R]esolution 1325 (2000) represented an enormous step forward in the protection of **women** and highlighted the importance of their role [...]"

---

[5] UNSC_2004_SPV.5066Resumption1_spch026

In (6), DeepSeekZero failed to annotate the entity at all, while DeepSeekFew even labeled “the armed forces” as "hero" despite their role in sexual violence:

(6) "[...] the adoption of a plan of action to combat sexual violence among **the armed forces** [...]"

The reasoning by DeepSeekFew offers an explanation for the misclassification:

> *"[...] the military justice system are Heroes as they’re part of the government’s efforts."*

It simply conflated the perpetrators as a part of the government with the heroic efforts of the government, overlooking an even explicit accusation. This example illustrates why relying solely on LLMs for tasks like character role identification can be dangerous.

While some models, like RoBERTa, generally perform well in terms of entity recognition, challenges remain in handling complex cases involving articles, pronouns, multi-word entities, and role shifts. The few-shot prompting approaches showed some improvement, but inconsistencies and biases still persist. These findings highlight the need for careful consideration of bias when applying LLMs to role classification tasks, particularly in contexts involving gender and institutional entities. Further refinements and research to determine where these biases originate from is crucial for improving their performance.

5 Conclusion

Inspired by the central question of how women are depicted in a discourse centered on them, this paper explored how entities are framed in UNSC debates. Using topic modeling, clustering, and manual annotations of 54 speeches, we found that women are the main characters and primary victims. However, they also appear as heroes, providing valuable knowledge and playing a significant role in peacekeeping. Frequent co-occurrence with "girls" and "children" reinforces their portrayal as innocent civilians, while they are never framed as villains.

Men appear far less often and are predominantly grouped with women and children, either as co-victims or co-heroes, highlighting the importance of cooperation between genders. Meanwhile, the UNSC often portrays itself as a hero by highlighting its contributions, though occasionally they acknowledge their own shortcomings. Apart from this, while some non-state actors are explicitly called out, perpetrators largely remain unnamed, and no Member State is explicitly accused. This reflects the diplomatic tone of UN discourse, where positive, optimistic visions are encouraged.

To assess the feasibility of automating role classification, we evaluated RoBERTa, and DeepSeek and Llama in zero- and few-shot settings. While RoBERTa performed reliably and sometimes even outperformed human annotations, the chat models often ignored key rules from the annotation guidelines and misclassified even obvious labels.

Despite limitations, our findings demonstrate the potential of framing analysis in institutional discourse and lay the groundwork for scalable, automated narrative role detection in international diplomatic settings, while emphasizing that human validation remains essential.

Future work could build on these insights by exploring country- and time-specific framing differences, addressing class imbalance through data augmentation, and experimenting with alternative model architectures. Additionally, further research is needed to identify and mitigate sources of bias in large language models to improve their effectiveness and fairness in this context.

Ethics Statement

This study uses publicly available political speeches from the United Nations Security Council and does not involve sensitive personal data. Three of the authors served as annotators; while manual annotation of character roles is inherently subjective, we attempted to mitigate bias through collaborative guideline development and consensus on labels. Given the highly subjective nature of the task, complete objectivity cannot be guaranteed.

We acknowledge that both the discourse analyzed and the computational models used may contain biases reflecting existing power structures and social inequalities. Since automated narrative role detection carries risks of oversimplification and misclassification, we emphasize that such methods should never replace, but rather complement, human interpretation.

All computational models used are publicly available, ensuring reproducibility and transparency.

Limitations

One of the primary limitations of this study is the scope of our annotations. Our dataset size was limited by time constraints, reducing model robustness. While the annotation process itself was not overly time-consuming, significant effort was required to define the annotation guidelines and curate the speeches collaboratively, making this stage particularly resource-intensive. A more extensive dataset would have provided additional training data and enabled a more robust evaluation of our approaches.

Moreover, seeing as the task was subjective, the annotators inevitably brought their own biases into the annotation process. This likely influenced both the evaluation and the model's performance, as the training data was shaped by our individual interpretations. These subjective factors further highlight the challenges of automating entity role classification and the need for continuous refinement of guidelines and methodologies. The subjective nature of detecting character frames is also reflected in the relatively low inter-annotator agreement. However, after extensive discussions, all three annotators reached consensus on the final labels, integrating different perspectives and thus improving data reliability.

Another limitation pertains to the computational and accessibility constraints associated with the models used. Many cutting-edge LLMs are not freely available and require subscription-based access, limiting their usability for broader research applications. Working locally restricted hyperparameter tuning; while our small dataset eased training, it also limits generalizability.

References

Çağlayan Başer. 2024. Women's role in violence and un women, peace, and security agenda. *Alternatif Politika*, 16(1):1–30.

Kelly Bergstrand and James M. Jasper. 2018. Villains, victims, and heroes in character theory and affect control theory. *Social Psychology Quarterly*, 81(3):228–247.

R. Charli Carpenter. 2005. "women, children and other vulnerable groups": Gender, strategic frames and the protection of civilians as a transnational issue. *International Studies Quarterly*, 49(2):295–334.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Stanley Cohen. 2011. *Folk devils and moral panics*. Routledge.

Victoria Martín de la Rosa and Luis Miguel Lázaro. 2019. How women are imagined through conceptual metaphors in united nations security council resolutions on women, peace and security. *Journal of Gender Studies*, 28(4):373–386.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Lea Frermann, Jiatong Li, Shima Khanehzar, and Gosia Mikolajczak. 2023. Conflicts, villains, resolutions: Towards models of narrative media framing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8712–8732, Toronto, Canada. Association for Computational Linguistics.

Kai Gehring and Matteo Grigoletto. 2023. Analyzing climate change policy narratives with the character-role narrative framework. Technical report, CESifo Working Paper.

Sheri Lynn Gibbings. 2011. No angry women at the united nations: Political dreams and the cultural politics of united nations security council resolution 1325. *International Feminist Journal of Politics*, 13(4):522–538.

Diego Gomez-Zara, Miriam Boon, and Larry Birnbaum. 2018. Who is the hero, the villain, and the victim? detection of roles in news articles using natural language techniques. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, pages 311–315.

Francesca Grasso, Stefano Locci, and Manfred Stede. 2025. Applying the character-role narrative framework with llms to investigate environmental narratives in scientific editorials and tweets. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 49–67.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

James M Jasper, Michael Young, and Elke Zuern. 2018. Character work in social movements. *Theory and Society*, 47:113–131.

Michael D Jones and Mark K McBeth. 2010. A narrative policy framework: Clear enough to be wrong? *Policy studies journal*, 38(2):329–353.

Stephen Karpman. 1968. Fairy tales and script drama analysis. *Transactional analysis bulletin*, 7(26):39–43.

Orrin E Klapp. 1954. Heroes, villains and fools, as agents of social control. *American sociological review*, 19(1):56–62.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart De Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th international conference on computational linguistics: system demonstrations*, pages 5–9.

Ludovic Kun, Jayesh Bankoti, and David Kiskovski. 2022. Logically at the constraint 2022: Multimodal role labelling. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 24–34, Dublin, Ireland. Association for Computational Linguistics.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–321, Berlin, Germany. Association for Computational Linguistics.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2329–2334.

Mirco Schoenfeld, Steffen Eckhard, Ronny Patz, Hilde van Meegdenburg, and Antonio Pires. 2019. The UN Security Council Debates.

Shivam Sharma, Tharun Suresh, Atharva Kulkarni, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Findings of the CONSTRAINT 2022 shared task on detecting the hero, the villain, and the victim in memes. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 1–11, Dublin, Ireland. Association for Computational Linguistics.

Pranaydeep Singh, Aaron Maladry, and Els Lefever. 2022. Combining language models and linguistic information to label entities in memes. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 35–42.

Dominik Stammbach, Maria Antoniak, and Elliott Ash. 2022. Heroes, villains, and victims, and GPT-3: Automated extraction of character roles without training data. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 47–56, Seattle, United States. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Jacqui True and Antje Wiener. 2019. Everyone wants (a) peace: the dynamics of rhetoric and practice on 'women, peace and security'. *International Affairs*, 95(3):553–574.

UNDP. 2023. Parliament as partners: Supporting the women, peace, and security agenda - a global handbook. Accessed: 2025-03-24.

Erika Allen Wolters, Michael D. Jones, and Kathryn L. Duvall. 2021. A narrative policy framework solution to understanding climate change framing research. *Narratives and the Policy Process : Applications of the Narrative Policy Framework*.

Yue Wu, Xuan Tang, Tom M. Mitchell, and Yuanzhi Li. 2024. Smartplay: A benchmark for llms as intelligent agents. *Preprint*, arXiv:2310.01557.

Appendix A UNSC Quote Filenames

Quote #	Filename
(1)	UNSC_2003_SPV.4852Resumption1_spch035.txt
(2)	UNSC_2019_SPV.8649_spch030.txt
(3)	UNSC_2015_SPV.7533_spch015.txt
(4)	UNSC_2013_SPV.6984_spch057.txt
(5)	UNSC_2011_SPV.6642_spch055.txt
(6)	UNSC_2016_SPV.7704_spch087.txt

Table 3: Corresponding filenames for numbered in-text examples.

Appendix B Interannotator Agreement

	Annotator 1	Annotator 2	Annotator 3	Curator
Annotator 1	–	0.30	0.44	0.47
Annotator 2		–	0.38	0.48
Annotator 3			–	0.70
Curator				–

Table 4: Krippendorff's Alpha (unitizing/character offsets) for agreement on role annotation (Hero, Victim, Villain) computed pairwise. Three of the authors served as annotators and worked collaboratively to define a gold standard, which is reflected in the Curator column.

Appendix C Heroes, Victims, and Perpetrators in WPS: Annotation Guidelines

i. Document Level Rules Exclude all president speeches

ii. Character Level Rules

a. We include entire noun phrases (NP): This includes any numerical or descriptive modifiers.
Example: "60 million Africans"
b. We annotate restrictive relative clauses fully.
Example: "parties that perpetuate acts of violence against women and children"
c. We include possessive modifiers within the NP:
Example: "my delegation"
d. When multiple characters are listed together as part of the same NP and share a common role or action, we consider the entire sequence as one entity, so we not split or annotate each item separately.
Example: "Allow me to begin by thanking **the Secretary-General, Mr. Kevin Hyland, Mr. Yury Fedotov and Ms. Ilwad Elman** for their briefings."
e. We do not include punctuation at the end of annotations.
Example: "**men and boys**."
f. We do not annotate predicate nominatives or any other descriptions of characters.
Example: "**Women** are not only victims." Here, 'victims' should not be annotated.
Example: "**women** as agents of peace"
g. We do not annotate positions or groups when they are only mentioned in the abstract, without references to the achievements of the particular entity.
Example: "the establishment of the post of the Special Representative of the Secretary General on sexual violence in situations of armed conflict"

Role-Definitions

In German, the constellation of hero, victim, and perpetrator is called the drama triangle, first introduced by Karpman (1968). It is a psychological model used in transactional analysis to describe interactions between individuals or groups as adopting one of three stereotypical roles: victim, perpetrator, or hero, which simplifies the linking of behaviors with stereotypical roles learned at an early stage (Karpman, 1968; Gomez-Zara et al., 2018). The link to stereotypes, the framing of the characters with said roles, influences the readers to have a certain bias and to develop a non-reflective opinion through the stereotypical knowledge, which is influenced by the author alone (Gomez-Zara et al., 2018). Hero, victim, and perpetrator are also of great importance in affect control theory and character theory (Bergstrand and Jasper, 2018). The following are definitions we will work with for the WPS-project:

Victim

Victims tend to be portrayed as weak, good, innocent people who are in need of protection (Jasper et al., 2018). Due to these characteristics, they often motivate and encourage action towards a specific cause and can help make aware of injustices which are worth com bating. Jasper et al. (2018) state that victim's sufferings are often elaborated in detail to arouse more moral emotions and indignation. "Popular" victims, as they get the most sympathetic reactions in the modern world due to their cultural innocence, are children (Jasper et al., 2018).

Task-specific additions (based on the WPS corpus):

- Someone who is excluded from decision-making processes/someone who is not given the recognition/power that they deserve.
- Someone who suffers acts of sexual violence/physical harm/displacement, etc.
- Someone who is not given the same equal rights as other parties

Perpetrator/Villain

Perpetrators are people whose moral reputation turns or has turned negative (Jasper et al., 2018). They are considered to be people who spread anxiety and fear, cause people to lose their daily routines, and make them sacrifice their lives, for example, within wars (Bergstrand and Jasper, 2018). Perpetrators often share the same characteristics as heroes, such as being strong, brave, and intelligent. However, their description tends to be more like that of beast-like predators: powerful, threatening, and delinquent (Cohen, 2011).

Task specific additions (based on the WPS corpus):

- Someone who is responsible for causing anxieties, damage, and crimes.
- Someone who stands in the way of equal rights and justice for victims.

Hero

Heroes are people who, by helping victims (and defeating the villains (Jasper et al., 2018; Klapp, 1954)), can become heroes. They are defined as people who increase agreement within groups and boost commitment to a cause. They tend to be well-intentioned people, who recognize injustice, try to resolve and fight it, as well as protect others. However, this does not mean that heroes are completely independent. Jasper et al. (2018) state that even a hero might be in need of help from an even more experienced hero. Furthermore, they are often put in the context of success (Klapp, 1954).

Task specific additions (based on the WPS corpus):

- Someone who has the potential to save others, given their abilities, knowledge, positionality/perspective, but who is not able to do so because of structural discrimination.
- Someone who makes a call to action
- Someone who is recognizing the unjust treatment and violation of victims or is calling it out.

Further Thoughts

It can be seen that the victim, hero, and perpetrator definitions align very much with stereotypes we are taught and confronted with since our childhood. However, roles are not fixed, as heroes can transform into victims or perpetrators (Jasper et al., 2018). The definitions are tried to be kept as simple and generally adaptive as possible, because depending on the source one works with, representations can change. Jasper et al. (2018), for example, state that in news articles, victims are kept nameless in contrast to heroes and perpetrators. In the WPS dataset, however, we observed that it is the perpetrators who most commonly remain unnamed.

Entities to tag

a. People
b. Organisations
c. Countries
d. Groups
e. UN Resolutions. Only annotate United Nations Resolutions if they are explicitly personified.

Example: In “Resolution 1325 calls for action,” the resolution should be tagged as a hero. In “working towards the implementation of resolution 1325,” the resolution should not be marked.

Entities NOT to tag

a. Abstract concepts: Abstract ideas or symbolic references.

Examples: “International cooperation,” “sexual violence,” “**women’s** participation.” In the latter example, women are the entities that are taking heroic action, while their participation refers to a concept.

b. Entities hoping for/welcoming/thanking/commending something good. These are passive actions.
c. Laws or Treaties.

Annotation of character role labels

We only annotate generic role terms (like “victim”) when no other specification is included.

Example: In “Victims of these atrocious crimes have been waiting for justice”, we would tag “victims of these atrocious crimes” as Victim. However, if specific entities are mentioned, like in “Victims of these atrocious crimes, namely women, have been waiting for justice” we only annotate “women” as Victim.

Annotation of multiple roles

In certain cases, characters may be portrayed with multiple roles simultaneously. When this happens, we annotate the entity with a combined role. Furthermore, the same entity can take on different roles throughout the speech. For instance, while “women” might be classified as "victim" in one sentence, the same entity can also be classified as "hero" in a different part of the speech.

Example: “The equal right to decision-making and participation, along with **women’s** empowerment, is crucial to ensure a functioning society and peace and justice in the aftermath of conflicts.”

Explanation: In this case, “women” are portrayed both as victims (since they need external help to be empowered) as well as heroes (since they contribute to a peaceful society). When an entity fits into multiple roles based on the context, we use combined tags.

iii. Annotation examples and explanations

Legend: Hero Perpetrator Victim

1. "We commend the work that has been done by the United Nations Children’s Fund in reintegration projects that has led to the release of girls from the armed forces in various countries."

In this example, "we" is not annotated as a hero because they are just thanking another organization.

However, the organization in question, the "United Nations Children's Fund," is marked as a hero because they are responsible for these projects that help girls. The "girls," on the other hand, are marked as victims because, although it could be argued that they are the benefactors of the situation (since they have been released), we believe that the fact that they were abducted in the first place carries more weight, also because they are in danger of being abducted again at any time. Most intuitively, "the armed forces" are the perpetrators, since they were responsible for the girls' abduction.

2. "Members of the Council note that although women have begun to play an important role in conflict resolution, peacekeeping and peace- building, they are still under-represented in decision-making in regard to conflict. If women are to play an equal part in security and maintaining peace, they must be empowered politically and economically, and represented adequately at all levels of decision-making, both at the pre-conflict stage and during hostilities, as well as at the point of peacekeeping, peace-building, reconciliation and reconstruction."

This example demonstrates that the same entity (in this case, "women") can take on multiple roles within one sentence. In the first instance ("women"), they are marked as heroes due to their important role in conflict resolution. However, in the second part of the sentence, where they are referred to as "they," the focus shifts to their underrepresentation in decision-making, marking them as victims. The conjunction "although" introduces this contrast, highlighting both aspects of the situation. In the second sentence, the focus remains on women as victims ("if women are to play an equal part in security"), implying that they are not currently able to do so. The need for empowerment, both politically and economically, is presented because they are not given equal opportunities as men. Therefore, they are portrayed as victims in this context.

3. "Those who commit crimes against women, including the peacekeeping personnel, should be brought to book. Let us heed the women's cry for an equal opportunity to voice their ideas in official peace negotiations. And let us act now."

In the last two sentences of this example, "us" is marked as a hero since the speaker is engaging in a call to action. The speaker portrays "us" as a hero by attempting to encourage others to acknowledge and act upon the unjust inequality faced by women.

4. "My Government attaches great importance to the protection and security of women and girls, both in situations of armed conflict and in peace."

In this case, "my government" is marked as a hero because, while the action is indirect, they recognize the importance of the issue and stand up for it. "Women and girls" are marked as victims since they require protection and security, as they cannot protect themselves (see victim definition above). Although they benefit from the protection measures (and thus could be seen as benefactors), we chose to prioritize their need for protection, which highlights their vulnerable position and role as victims in society.

5. "Being direct victims of violence and discrimination, women have gained a great understanding of the need to address peace comprehensively."

Here, "women" are simultaneously considered heroes and victims. The victim annotation is indicated by the context of them being the "direct victim," while the hero annotation relies on the positionality of women–their potential to provide valuable knowledge.

Appendix D RoBERTa Model

Appendix D.1 BIO-Annotation Example

Token	"[...]	my	country	has	[...]	defending	the	victims	of	terrorism	."
Labels	O	HERO	HERO	O	O	O	VICTIM	VICTIM	VICTIM	VICTIM	O
BIO Labels	O	B-HERO	I-HERO	O	O	O	B-VICTIM	I-VICTIM	I-VICTIM	I-VICTIM	O

Table 5: Example of the BIO annotation scheme applied to a sample sentence from `UNSC_2015_SPV.7585_spch010.txt`. The whole sentence is: *Council members know that my country has been a standard-bearing country for defending the victims of terrorism.* Entity spans are labeled as "B-" (begin), "I-" (inside), or "O" (outside)

Appendix D.2 Hyperparameter Optimization and Model Selection

During fine-tuning we limited the amount of values to ensure computational feasibility within the available resources.

	Learning Rate	**Epochs**	**Optimizer**
m1	5e-5	3	AdamW
m2	5e-5	30	AdamW
m3	1e-5	3	AdamW
m4	1e-5	30	AdamW
m5	5e-5	30	SGD
m6	3e-5	30	SGD

Table 6: Hyperparameters for m1-m6

The performance of the fine-tuned models on the development dataset is shown in Table 7, based on their F1 values for models m1-m4. Models m5 and m6 are excluded from the table as their results were identical to those of m4.

	F1				**S**
	m1	m2	m3	m4	
HERO	0.59	0.62	0.63	**0.64**	122
VICTIM	0.67	0.72	**0.73**	**0.73**	77
VILLAIN	**0.54**	0.52	0.5	0.52	10
HERO_VICTIM	0.45	**0.46**	0.36	0.4	26
HERO_VILLAIN	0	0	0	0	0
VICTIM_VILLAIN	0	0	0	0	3
micro avg	0.6	0.62	0.62	**0.64**	238
macro avg	0.45	**0.46**	0.44	0.38	238
weighted avg	0.59	0.62	0.62	**0.63**	238

Table 7: F1 scores of the models m1, m2, m3, and m4. S (support) represents the number of instances per class in the dataset.

The results show that models m3 and m4 achieve the highest F1 scores for the two most frequently occurring classes, heroes (m3: 0.63; m4: 0.64) and victims (m3 & m4: 0.73). m4 yields the highest weighted average F1 score (0.63). These findings served as the primary rationale for selecting the hyperparameters of m4 for the final fine-tuning process. The final model was fine-tuned using both the train and the dev set.

Appendix E Figures

Appendix E.1 Annotation Insights

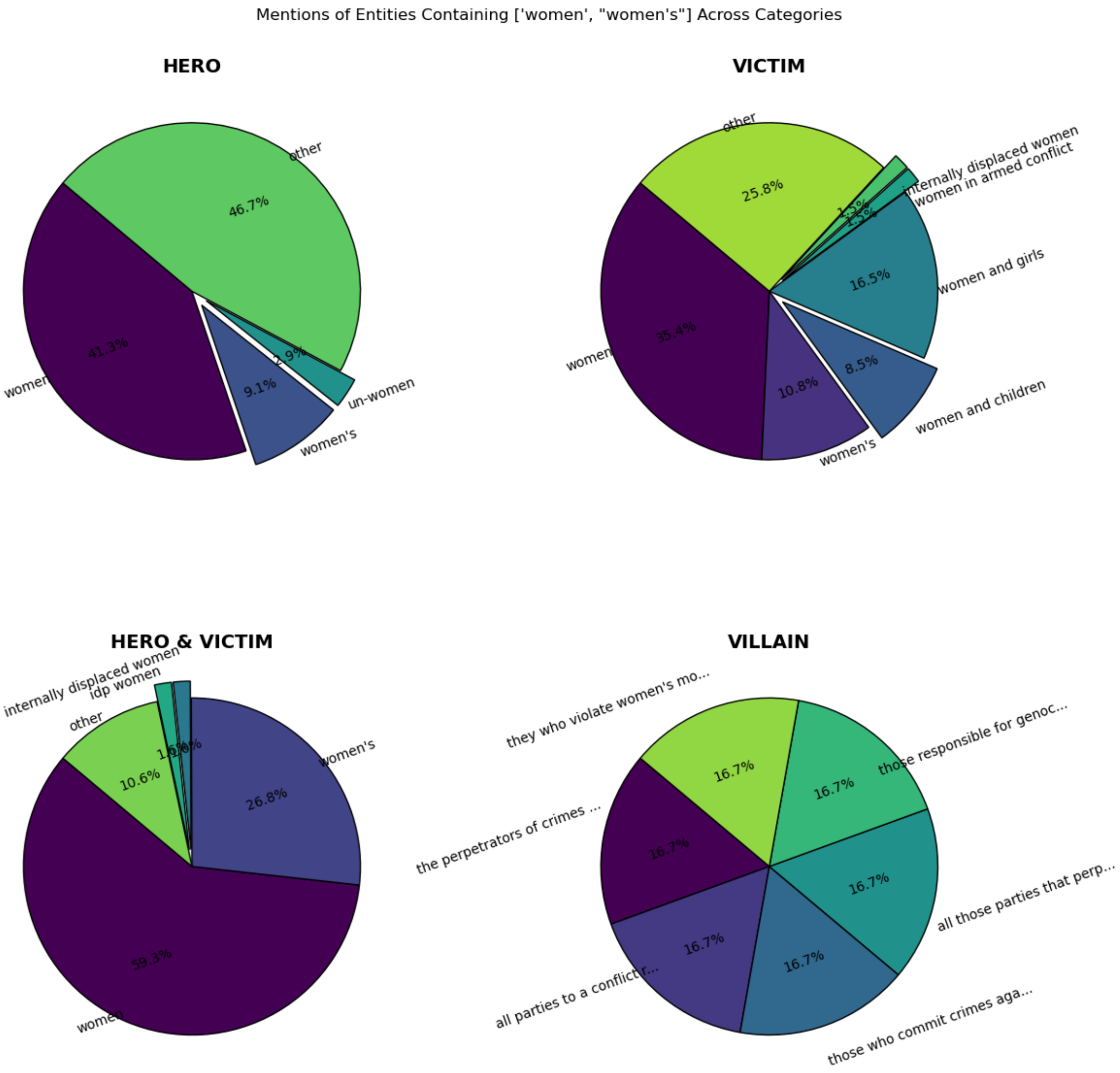


Figure 4: Distribution of Characters for Women Entities

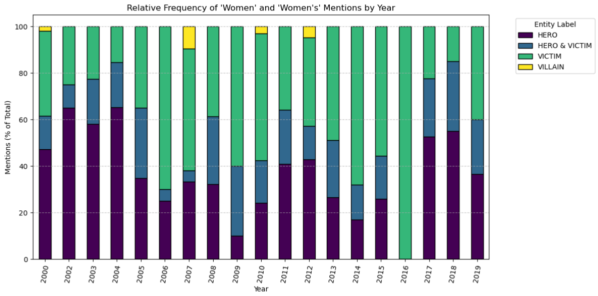


Figure 5: Character Labels for "women" and "women’s" Across the Years

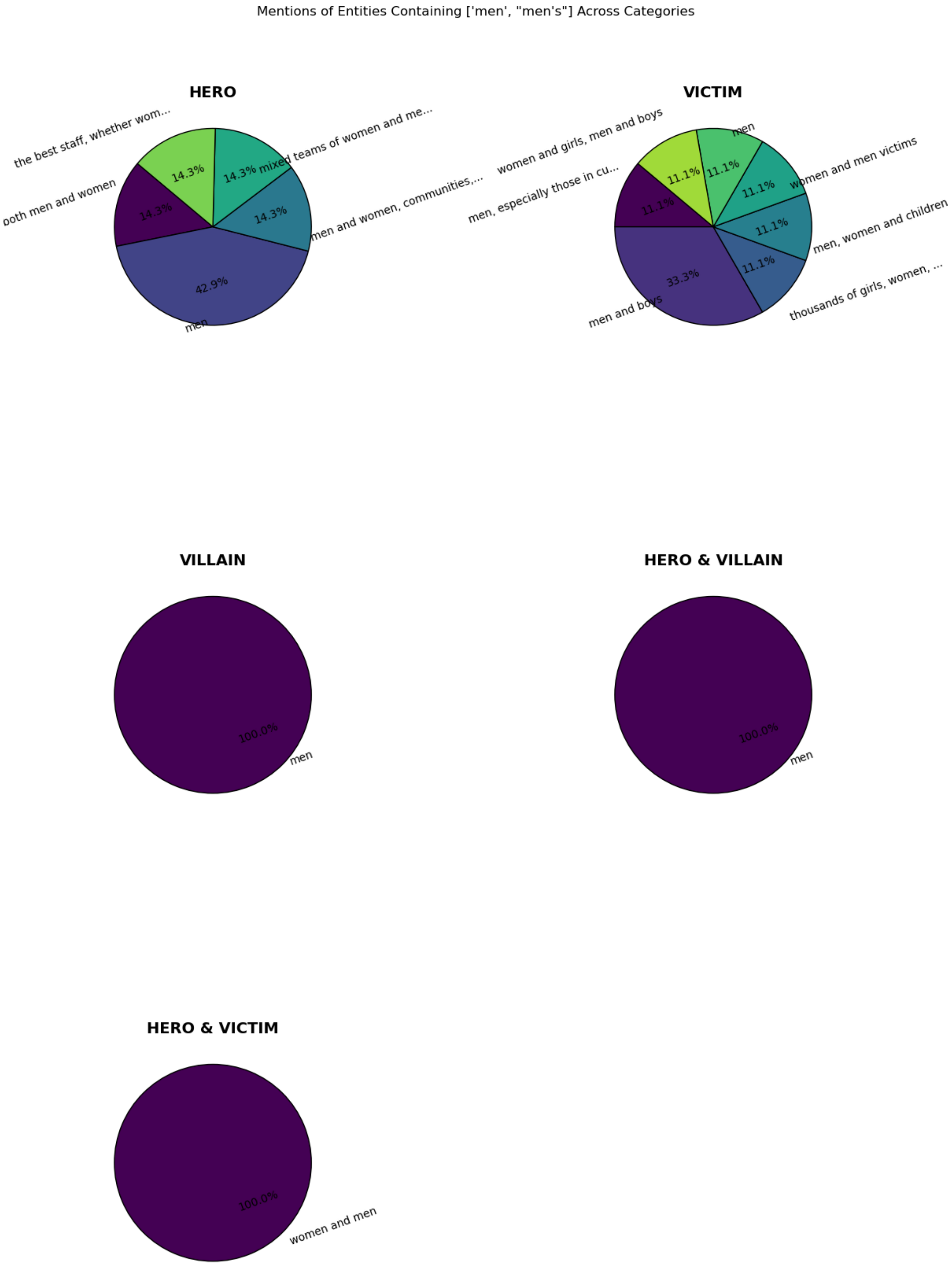


Figure 6: Distribution of Characters for Men Entities

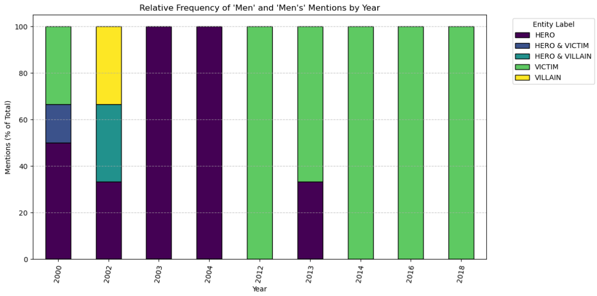


Figure 7: Character Labels for "men" and "men’s" Across the Years

Appendix E.2 Meta-Analysis

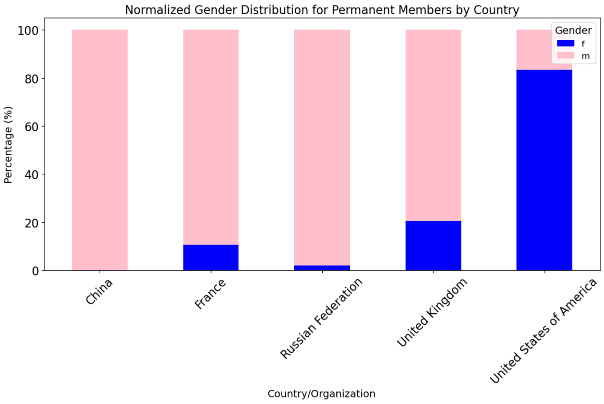
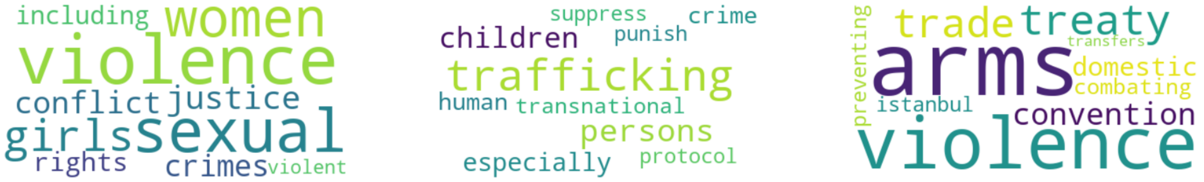


Figure 8: Gender Distribution of Speakers Among Permanent Members of the UNSC

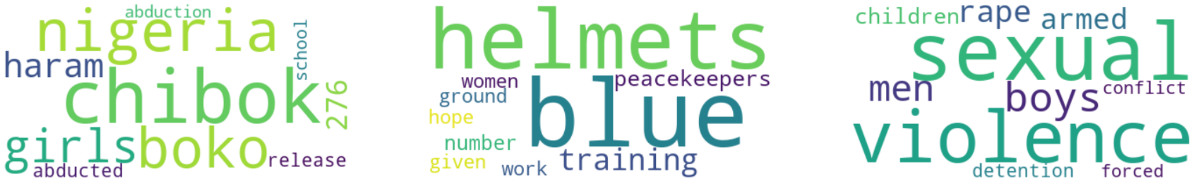
Appendix E.3 BERTopic



(a) Female Entities: Sexual violence

(b) Female Entities: Human trafficking

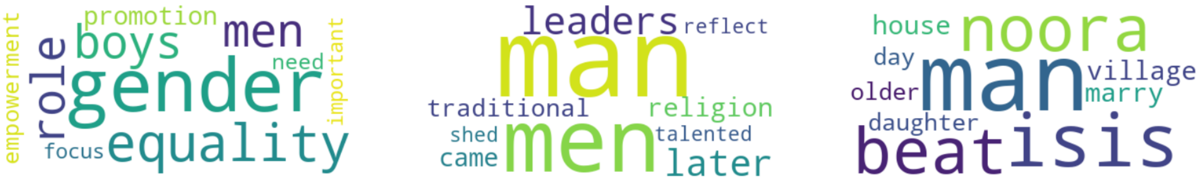
(c) Female Entities: Domestic Violence



(d) Female Entities: Terror Organizations

(e) Female Entities: Blue Helmets

(f) Male Entities: Sexual violence



(g) Male Entities: Promotion of Gender Equality

(h) Male Entities: Overlapping topics

(i) Male Entities: Victim's reports

Figure 9: BERTopic visualizations for male and female entities.

Appendix E.4 Clustering

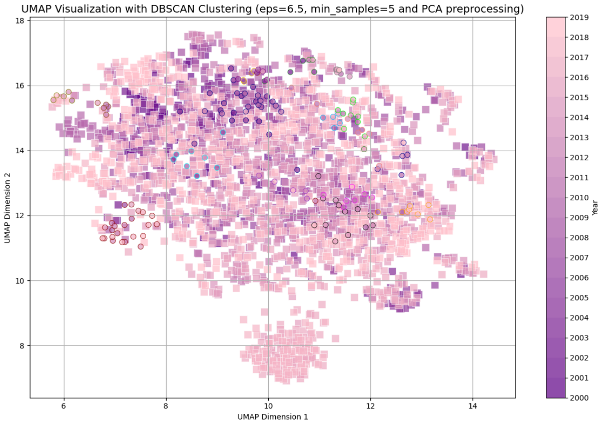


Figure 10: Clusters as identified using DBSCAN. Each point represents a speech. Points that are square-shaped with a white outline represent outliers, while circle-shaped points represent speeches that are part of different clusters, distinguished by the color of the outline. The fill color of each point represents the year that the speech was held.

Enhancing Policy Analysis with NLP: A Reproducible Approach to Incentive Classification

M.A. Waskow and **John P. McCrae**
Data Science Institute
University of Galway
Insight Research Ireland Centre for Data Analytics

Abstract

Over the past few decades, political scientists have increasingly used Natural Language Processing (NLP) methods in their research. Within the subdomain of public administration, there remain further opportunities for the use of NLP in the task of policy analysis. The potential of a machine learning pipeline to identify sentences containing incentives has been demonstrated for the Spanish-language forestry policies of five Latin American countries, but the study was not reproducible due to a lack of model and data availability. This paper seeks to validate the existing pipeline of policy collection, sentence labelling, fine-tuning, and incentive classification by replicating it in a new context and achieving comparable performance, as well as to publish all relevant data and training information to ensure reproducibility. In the domains of a different language and geopolitical system, namely English-language Irish forestry policies, this implementation demonstrates the pipeline's transferability by achieving mean overall F1 scores of 88.3 for binary classification and 96.8 for multiclass classification with our best models. The contributions of this paper are twofold: the validation of an existing pipeline by replicating it in new geopolitical and linguistic domains, and the creation of a novel open dataset of Irish forestry policy sentences labelled for incentive classification.

1 Introduction

Political and social scientists have been utilizing Natural Language Processing (NLP) methods in their research for decades in order to better analyse text as data (Laver et al., 2003; Grimmer and Stewart, 2013). NLP-based analyses of political communication texts — including speeches, manifestos, news articles, and tweets — have provided unique insights into political parties, propaganda, and public opinion. While the field of public policy has been slower in implementing NLP, researchers are beginning to explore its potential applications in the area of policy analysis.

A core task of policy analysis is the development of reports to inform the creation or maintenance of policies (i.e. official proposals of actions to be taken by a government or institution) by presenting evidence-based solutions to a given problem. The research behind these reports can help prevent policy failure by combatting the inappropriate solutions and ambiguity that threaten successful policy implementation, but the thoroughness of policy analysis requires a high time cost that an analyst cannot always afford. NLP can help to reduce this bottleneck of policy creation, making the process more efficient and leading to earlier real-world changes.

Firebanks-Quevedo et al. (2022) demonstrated that their NLP pipeline could help streamline key tasks of policy analysis: finding policies about a given topic and identifying the policy instruments used within them. Policy instruments are methods of government intervention to achieve certain outcomes, like laws (regulation), grants (economic action), or propaganda (communication); this project focused specifically on economic instruments called incentives, which encourage actions and behaviours through financial rewards or pressures like subsidies or taxes (Badie et al., 2011). The pipeline consists of scraping policies from the websites of five Latin American governments, extracting and labelling their sentences for presence and type of policy incentive, then training both a binary and a multiclass classifier on the labelled dataset. The original dataset and models are not available, however, limiting the reproducibility as well as the implementation of their work.

Despite Dodge et al. (2019) offering a structured checklist to address the reproducibility crisis in NLP, many projects do not make core information relating to their research and the production of their results available. This challenges not only the cred-

Proceedings of the 5th Workshop on Computational Linguistics for the Political and Social Sciences (CPSS), pages 68–79
September 10-11, 2025 ©2025 Association for Computational Linguistics

ibility of the conclusions, but also the feasibility of downstream impacts of their findings. Magnusson et al. (2023) has shown that even a few years after the publication of the checklist, the issues of missing data, code, and model training information are still ongoing in NLP.

This project replicates the pipeline of Firebanks-Quevedo et al. (2022) in a new domain, also addressing its reproducibility issues through the publication of our novel dataset[1] and updated code[2]. Where the original workflow was based on Spanish-language texts from Latin American forestry policies, we focus on the English-language texts of Irish forestry policies. We maintained the policy domain of forestry as our implementation already introduced changes to the language and country of interest. Additionally, we further updated the pipeline with new models and accommodations for imbalanced datasets to improve performance. Our final results validate the original pipeline by demonstrating its transferability across linguistic and geopolitical contexts, as well as contribute a new resource for NLP and policy analysis through the release of our code and 1.4k sentence dataset labelled for binary and multiclass incentive classification.

2 Related Work

NLP has become a familiar tool for political and social scientists working with diverse and ever-growing text datasets for diverse and ever-growing applications. For news analysis, classifiers have been trained to identify different topics or agendas contained in articles, even so far as to identify the use of different propaganda techniques (Yoosuf and Yang, 2019; Terechshenko et al., 2020; Nelson et al., 2021). Studying political parties and their messaging, classification and topic clustering have provided novel insights into the speeches of politicians and their electoral manifestos (Glavaš et al., 2017; Wilkerson and Casas, 2017; Rheault and Cochrane, 2020). Considering public opinion, Hagen et al. (2015) used topic modelling to explore policy suggestions from public petitions, while Terechshenko et al. (2020) performed sentiment analysis on tweets about an electoral candidate. Across these various political media, NLP has successfully helped scientists to synthesise new information and extract hidden patterns from the often-unstructured text data.

Progressing more into the subdomain of public administration and policy, Żółkowski et al. (2022) applied topic modelling and clustering to explore how EU countries were framing their climate policies, and Ningpeng et al. (2024) performed topic mining and text parsing on Chinese financial policy documents. Brandt (2019) additionally examined the restoration policies of three East African countries through paragraph topic classification. While these projects provided insight into the priorities of policies, their outcomes appear to be aimed more at political scientists than specifically policy analysts. In contrast, the sentence-level incentive classification of Latin American forestry policies in Firebanks-Quevedo et al. (2022) was explicitly aimed at aiding the task of policy analysis, though the inaccessibility of the models and dataset complicated its impact. Finally, Sewerin et al. (2023) have created a policy design annotations (POLIANNA) dataset, consisting of labelled text spans from EU climate policies and legal documents to provide structured data for future policy analysis tools. In summary, the application of NLP methods to analyse policy documents has largely followed the trend of topic modelling, clustering, and classification used on other political texts, but this is beginning to shift towards the use of sentence classification on policy documents to aid in the task of policy analysis.

3 Methodology

We gathered our policies by scraping them from the website of the Irish government, processed the PDFs into sentences, cultivated a dataset through both manual and human-in-the-loop (HITL) labelling processes, then trained and evaluated both the binary and multiclass classifiers. Figure 1 demonstrates this process.

3.1 Data Collection and Preprocessing

We began looking for Irish forestry policies in a repository of policies on the website of the Irish government[3]. At the time of our search, this page was still in development. The structure of the page was a list of the Irish government's departments and major initiatives wherein which a user, upon selecting a title in the list, would be taken to that topic's page. Then, on that topic's page, a user

---

[1] `https://huggingface.co/datasets/mawaskow/irish_forestry_incentives`

[2] `https://github.com/mawaskow/policy-classifier`

[3] `https://www.gov.ie/policies/`

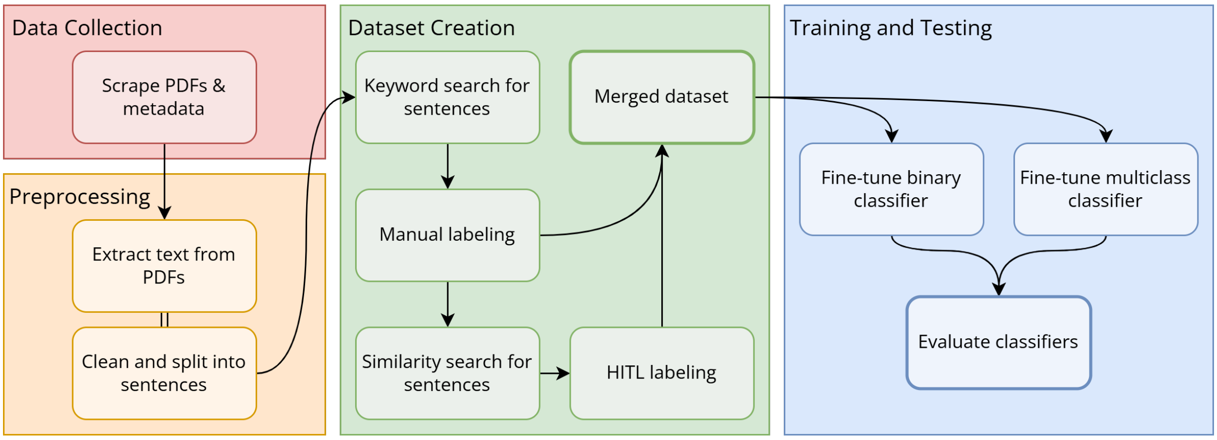


Figure 1: The incentive classifier pipeline

could find a link to a page where they could finally search that topic's publications, including policies. For a human, trying to find all the relevant policies across all the topics pages would be tedious and time-consuming, so we built a web crawler to help us scrape the data we needed.

We used Python's Scrapy[4] library to create a new "spider" or web crawler to extract the Uniform Resource Locators (URLs) for all the topic publication search pages. With this list, we constructed search query strings using keywords[5] for policy inclusion based on those used by Firebanks-Quevedo et al. (2022). Our crawler then visited all the search result pages and parsed through the queried policies to gather the policy metadata and document URLs, using exclusion keywords based on those in the original pipeline to ignore irrelevant policies and documents. In all, our scraper collected 138 relevant forestry policy documents across all the topic policy lists. A breakdown of the policies by source department can be found in Appendix A.

Once we had the collection of PDFs, we updated some of the preprocessing scripts of Firebanks-Quevedo et al. (2022) to retrieve the text from the PDFs using Python's PyPDF[6] library, which is based on text extraction instead of Optical Character Recognition (OCR). We then cleaned the texts of excess tags, spaces, and URLs and split them into sentences for labelling using a sentence tokeniser from Python's NLTK[7] library.

3.2 Dataset Creation

We follow the dataset creation steps of Firebanks-Quevedo et al. (2022), first hand-labelling a subset of the collected sentences, then using those labelled sentences as the basis for a HITL process to automatically label remaining sentences. One labeller completed both the manual and HITL steps, then a policy researcher annotated a stratified subset of the data in order to validate the labelling.

We classified sentences into one of six incentive classes, or as a non-incentive, as defined in the reference pipeline:

- **Credit**: Loans, insurance
- **Direct payment**: Cash, grants
- **Fine**: Penalty payment
- **Supplies**: Material support, equipment
- **Tax deduction**: Reduced tax liability
- **Technical assistance**: Training, experts

Firebanks-Quevedo et al. (2022) had noted that their pipeline was unable to distinguish between intentions, plans, or general mentions of incentives in the sentences. In an attempt to address this, we labelled sentences that simply mentioned incentive keywords like *grant* or *loan* as a non-incentive, while sentences which actually declared the creation or implementation of incentives were classified by incentive type.

3.2.1 Manual Labelling

For our first pass at labelling the sentences, we performed a keyword search of incentive-related

---

[4] `https://www.scrapy.org/`

[5] `https://github.com/mawaskow/policy-classifier/blob/main/policy_scraping/policy_scraping/keywords/keywords_forestry.json`

[6] `https://pypdf.readthedocs.io/`

[7] `https://www.nltk.org/`

substrings, then manually assigned labels to a subset of the possible incentive sentences. The hand-labelling resulted in 965 sentences across the seven classes.

3.2.2 HITL Labelling

Firebanks-Quevedo et al. (2022) conducted their HITL labelling by using five sentences for each incentive class as queries to perform a similarity search in the remaining sentences. Of the sentences returned, the original pipeline only kept sentences that occurred across all five queries for each class. As our data was reflective of one country instead of five, the size of our dataset required that we lower the inclusion criteria to occurrence within at least four of the five queries for each label. Otherwise, we followed the same steps and produced an additional 626 pre-labelled sentences across the seven classes, subsequently validated by our labeller.

3.2.3 Novel Dataset

Our final dataset of the incentive sentences was the result of merging the manually labelled and HITL-labelled datasets. In order to account for duplicate sentences across the two collections, we grouped the sentences with Levenshtein or edit distances above the hand-tuned threshold of 0.9 and removed redundant entries. The final dataset consisted of 1419 labelled, filtered sentences containing both binary and multiclass labels.

The dataset overall had a notable but expected skew towards non-incentives, with incentives making up only 18.5% of the dataset. The classes of incentives were also unbalanced themselves, with Supplies at 30.1% of all incentive sentences, Technical assistance at 27.9%, Direct payment at 23.0%, then Fine at 8.6%, Credit at 7.1%, and Tax deduction at 3.3%.

For validation of our labelling, we brought in a policy researcher to label a stratified subsample of our final dataset. They annotated 10.7% of the sentences, including 24.5% of the incentives. To evaluate our agreement, we calculated the Cohen's kappa score of the resulting binary and multiclass datasets (Cohen, 1960). The agreement of the binary labelling, computed across the entire dataset, was 0.631, considered "substantial agreement." For the multiclass labelling, we evaluated agreement across the subset of examples where both annotators labelled the sentence as some kind of incentive, resulting in a score of 0.859, "near perfect agreement." Further information about validator agreement can be found in Appendix B.

Parameter	Binary	Multiclass
Epochs	5	15
Batch Size	16	16
Learning Rate	2E-5	2E-5
Weight Decay	0.01	0.01
Optimiser	AdamW	AdamW

Table 1: Hyperparameters for binary and multiclass model training

3.3 Training and Testing

The next step in the pipeline is the fine-tuning of models on each of the binary and multiclass versions of the dataset. All training was conducted on a CUDA-enabled NVIDIA GeForce RTX 3080 Laptop GPU with 16 GB VRAM, appropriate for small to medium size models.

To produce the binary dataset, we used all 1419 sentences, keeping the non-incentive label and replacing all incentive class labels with "Incentive." The multiclass dataset consisted of the 263 incentive sentences from the final dataset with no additional alterations.

Due to the small size and significant class imbalances of the datasets, we chose to train and evaluate our classifiers across ten random train-dev-test splits (60/20/20) of each of the binary and multiclass datasets, stratified to maintain label proportions. To ensure reproducibility and controlled randomness, each split was generated using a distinct random seed (ranging from 0 to 9). We averaged our final metrics across all ten runs to account for variations in performance across different splits, providing a more robust estimate of model performance.

3.3.1 Model Selection

We first established the baseline for our replication by using the same sentence-transformer (Reimers and Gurevych, 2019) model and hyperparameters as the original pipeline for our binary and multiclass classification. The reference model, sentence-transformers/paraphrase-xlm-r-multilingual-v1 (XLM-R), is a multilingual transformer with 278M parameters (Conneau et al., 2020).

Following the first model, we explored whether newer or more efficient models could improve performance. We tried another multilingual SBERT model with the same number

Model		Binary	Multiclass
XLM-R	d	87.7 ± 1.4	94.1 ± 4.2
	l	87.5 ± 1.9	94.3 ± 3.5
	o	87.5 ± 1.3	94.7 ± 3.7
MPNet	d	**88.3 ± 1.5**	95.7 ± 2.5
	l	88.0 ± 2.3	96.2 ± 2.5
	o	88.2 ± 1.0	**96.8 ± 2.1**
GTE	d	87.1 ± 1.1	95.6 ± 3.2
	l	87.2 ± 1.4	95.3 ± 2.7
	o	87.7 ± 1.2	96.3 ± 2.7
E5	d	87.4 ± 1.6	95.7 ± 2.9
	l	86.8 ± 1.7	95.8 ± 2.0
	o	86.5 ± 2.2	94.9 ± 3.3

Table 2: Overall F1 scores of the fine-tuned models (d: default, l: weighted loss, o: oversampling) averaged across 10 different dataset splits

Model		Non-Incentive	Incentive
XLM-R	d	92.5 ± 1.0	66.7 ± 3.6
	l	92.4 ± 1.2	66.1 ± 4.9
	o	92.7 ± 0.8	64.7 ± 3.8
MPNet	d	93.0 ± 0.9	**68.1 ± 4.4**
	l	92.8 ± 1.4	67.0 ± 6.1
	o	**93.2 ± 0.6**	66.6 ± 2.9
GTE	d	92.2 ± 0.8	64.8 ± 2.5
	l	92.4 ± 0.9	64.9 ± 3.9
	o	92.7 ± 0.8	66.0 ± 3.5
E5	d	92.5 ± 1.0	65.3 ± 4.3
	l	92.0 ± 1.1	63.9 ± 4.3
	o	92.1 ± 1.3	62.1 ± 6.2

Table 3: Label-specific F1 scores for the binary dataset (d: default, l: weighted loss, o: oversampling) averaged across 10 different dataset splits

of parameters, sentence-transformers/paraphrase-multilingual-mpnet-base-v2 (MPNet), which offers improved context capturing through its blending of permutation language modelling (PLM) with masked language modelling (MLM) (Song et al., 2020). We additionally tried two more recent, lightweight (109M parameters), English-only models, thenlper/gte-base (GTE) (Wang et al., 2022) and intfloat/e5-base-v2 (E5) (Li et al., 2023), to examine the possibility of pipeline deployment in environments with fewer computational resources–common in the political and social sciences. To ensure controlled comparisons across models, we used the same hyperparameters for all experiments as shown in Table 1.

Beyond updating the models, we also attempted to address the class imbalances of the dataset through two common strategies. In our first approach, we weighted the cross entropy loss, modifying the loss function to more severely penalise misclassification of the underrepresented classes during training; our class weights were inversely proportional to the class frequencies. Our other approach was to automatically oversample the minority classes with Imbalanced-learn's[8] RandomOverSampler, reinforcing the model's exposure to the underrepresented classes by balancing their distribution in the training data of each split.

3.3.2 Classification

Firebanks-Quevedo et al. (2022) determined that their pipeline performed best when they used the fine-tuned models to generate embeddings of the dataset, then sent those embeddings to a support vector machine (SVM) for classification (Gunn, 1998). Before we conducted the classification this way, we explored the inference capabilities of the fine-tuned transformer heads but found that performance on our small, imbalanced dataset was consistently higher when using the external classifier. For that reason and to further support our replication objective, all results reported are from the embedding generation and SVM classification method.

4 Results

We examined the F1 scores for the SVM classification of each model's sentence embeddings, averaged across the ten splits of the binary and multiclass datasets.

The mean F1 score and standard deviation for all models across the default, weighted loss, and oversampling training runs can be found in Table 2. MPNet outperformed the XLM-R model used in the reference pipeline and both other models, reporting a best average F1 score of 88.3 for binary classification and 96.8 for multiclass classification.

Overall, for binary classification, MPNet's embeddings achieved the highest average F1 scores, reporting a best value of 88.3. MPNet's performance was then followed by the XLM-R model, with the GTE and E5 models performing similarly but slightly worse than XLM-R. For multiclass classification, the MPNet again achieved the highest average F1 scores, this time reporting a best value of 96.8. After the MPNet, the rankings followed as

---

[8] `https://imbalanced-learn.org/stable/`

Model		Credit	Direct Payment	Fine	Supplies	Tax Deduction	Technical Assistance
XLM-R	d	97.1 ± 5.7	93.3 ± 5.1	93.3 ± 7.9	95.7 ± 4.0	81.7 ± 32.0	93.4 ± 4.8
	l	97.1 ± 5.7	92.1 ± 6.1	91.7 ± 9.6	96.5 ± 3.9	81.7 ± 32.0	94.7 ± 3.7
	o	**98.6 ± 4.3**	94.3 ± 5.4	94.4 ± 8.0	96.2 ± 3.7	78.3 ± 31.7	93.5 ± 3.7
MPNet	d	94.3 ± 7.0	93.5 ± 4.2	97.5 ± 7.5	97.1 ± 3.1	95.0 ± 15.0	95.7 ± 2.9
	l	93.2 ± 8.8	94.0 ± 3.9	97.8 ± 4.4	**98.4 ± 2.1**	96.7 ± 10.0	95.8 ± 4.2
	o	94.6 ± 8.6	**95.5 ± 2.8**	98.9 ± 3.3	97.5 ± 2.3	96.7 ± 10.0	**96.9 ± 2.9**
GTE	d	94.6 ± 6.7	93.7 ± 4.6	**100.0 ± 0.0**	96.5 ± 2.1	90.0 ± 30.0	95.3 ± 3.7
	l	89.8 ± 10.1	94.4 ± 3.5	**100.0 ± 0.0**	97.2 ± 2.6	90.0 ± 30.0	94.3 ± 3.6
	o	97.1 ± 5.7	95.3 ± 4.1	97.5 ± 7.5	97.2 ± 2.6	85.0 ± 32.0	96.4 ± 2.3
E5	d	94.6 ± 6.7	94.5 ± 5.0	97.8 ± 4.4	97.2 ± 2.5	96.7 ± 10.0	94.5 ± 4.4
	l	93.2 ± 6.9	94.6 ± 3.1	**100.0 ± 0.0**	96.9 ± 3.4	**100.0 ± 0.0**	94.7 ± 2.6
	o	95.7 ± 6.5	93.2 ± 4.7	97.5 ± 7.5	96.5 ± 3.1	95.0 ± 15.0	93.6 ± 4.4

Table 4: Label-specific F1 scores for the multiclass dataset (d: default, l: weighted loss, o: oversampling) averaged across 10 different dataset splits

GTE, then E5, then the XLM-R.

Tables 3 and 4 include the F1 scores for each classification label in the binary and multiclass datasets, respectively. Our best binary model's embeddings achieved average F1 scores of 93.0 and 68.1 for the classes of Non-incentive and Incentive, and our best multiclass classification model achieved average F1 scores of 94.6 for Credit, 95.5 for Direct payment, 98.9 for Fine, 97.5 for Supplies, 96.7 for Tax deduction, and 96.9 for Technical assistance.

We additionally report the overall validation F1 scores of our models in Appendix C as recommended by Dodge et al. (2019), as well as the overall and label-specific precision and recall scores of our models in Appendix D.

4.1 Qualitative Analysis

Examples of accurately and inaccurately classified binary incentive sentences can be found in Table 5. These sentences, embedded by MPNet and classified via SVM, highlight the semantic challenges of incentive sentence labelling and classification.

In the successfully identified examples, the label of the incentive sentence is clear due to its incentive-related keywords of *financial support*, *payments*, and *beneficiaries*, numerical amounts, and use of *will* to denote action. While the non-incentive sentence is a bit ambiguous due to the keywords *taxation* and *incentivising*, the use of *committed* demonstrates that this is a statement of intention or policy aspiration instead of a properly declared incentive, likely leading to its correct classification.

In the unsuccessful classification examples, the incentive sentence states the requirement in a scheme to apply financial penalties to herdowners but was marked as a non-incentive. Despite the use of keywords, this type of incentive– fine– may be better understood as a disincentive, so its lack of mention of rewards as found in other incentive sentences may have contributed to its misclassification. The non-incentive sentence mistakenly classified as an incentive did mention *direct payments*, but was in reality a description of a mechanism currently in place. The model likely picked up on the incentive keywords but missed the overall context of being a factual statement of existing supports rather than being an actionable incentive.

5 Discussion

Following the identification of incentive sentences through binary classification, the multiclass classification of incentive type is able to achieve high performance. In the binary context, we faced a challenge also noted by Firebanks-Quevedo et al. (2022), that it is difficult to distinguish incentive declarations from non-incentive sentences that mention incentives or state intentions to create them. At the label-specific level, the performance of incentive classification was worse than non-incentive classification, assumedly due to the imbalance of the dataset.

Regarding the different training methods, in the binary classification setting, the loss and oversampling methods did not appear to improve on default model performance, sometimes worsening it. In the multiclass classification setting, however, XLM-R, MPNet, and GTE models all benefited from the introduction of oversampling to the training.

The multiclass labels of Credit and Tax deduction received notably higher standard deviations across models and methods, though in the case of Tax deduction, this standard deviation was reduced through updating the embedding model to MPNet as well as by adding weighted cross entropy loss or oversampling in the training process. The high standard deviations were likely due to these being the most under-represented incentive classes in the dataset, resulting in high variability of the success of their classification across training splits. The challenge of correctly classifying the Tax deduction label was also interestingly consistent with the results of Firebanks-Quevedo et al. (2022).

As our best-performing model for generating embeddings was a multilingual MPNet sentence transformer, we share in the hope of Firebanks-Quevedo et al. (2022) that the models from our pipeline can be used to classify the sentences of policies in another language via cross-lingual transfer learning, removing the need to construct a whole new dataset for fine-tuning.

While the outcomes of this research primarily serve the development and evaluation of NLP applications for incentive sentence classification in policy texts, the practical impact on policy analysts remains indirect at this stage. As most policy analysts are not comfortable with building their own NLP pipeline implementations, the immediate application of this dataset and models is limited. However, we envision future work that translates these models into accessible tools, namely an interface where policy analysts can input search terms and websites of policy repositories to automatically retrieve relevant documents, or where analysts can upload policy documents to extract incentive sentences using the classification pipeline– first identifying incentives with our binary classifier, then categorising them with our muticlass classifier. We hope to soon make our pipeline accessible in this way to bridge the gap to impact real-world policy analysis, with the ultimate goal of improving the policymaking process and preventing policy failure.

6 Conclusion

This paper applied an existing policy incentive classification pipeline to a new geopolitical and linguistic context, demonstrating the transferability of the reference pipeline and creating a novel dataset of Irish forestry policy sentences labelled for incentive classification. The binary and multiclass classification of sentence embeddings produced by our best models achieved similar performance to the original Firebanks-Quevedo et al. (2022) pipeline's results on their own policy dataset, serving as a validation of their methodology.

We prioritised reproducibility in this replication, and encourage researchers to implement this or similar pipelines for policy incentive classification in more domains across languages, political contexts, and policy areas. Additionally, our dataset and training information is available for anyone who wants to fine-tune their own multilingual incentive classification models on an existing dataset for transfer learning into new contexts.

We hope that with more progress in the area of automatic policy incentive or instrument classification, this work can help streamline the task of policy analysis to enable robust recommendations of policy solutions, ultimately working towards the creation of more successful policies.

Limitations

Our pipeline did encounter limitations, some of which were shared with Firebanks-Quevedo et al. (2022) and some of which were unique to this implementation. We encountered the same challenge that Firebanks-Quevedo et al. (2022) did concerning the ambiguity of incentive sentences across intentions, plans, and mentions, resulting in suboptimal incentive identification performance in the binary classification context. Further examples of ambiguous sentences are presented in Appendix E.

Specific to this pipeline, in our labelling process we found that there were no incentive sentences about providing direct material support or equipment, so we adapted the Supplies definition to include grants which were exclusively for the purchase of materials or equipment. Additionally, when it came to labelling sentences about carbon taxes, we were conflicted between labelling it a Tax deduction since it is a tax mechanism, or including it with Fine since it can be considered a penalty or disincentive; we decided to proceed with the latter option. Additionally, the training and testing of our

Successes	True Label
Financial support towards the professional costs, such as legal, taxation and advisory for older farmers will contribute 50% of such vouched costs, to a maximum payment of €1,500 per beneficiary.	Incentive
We are committed to further developing a taxation framework, which plays its full part in incentivising, along with other available policy levers, the necessary actions to reduce our emissions.	Non-incentive

Errors	True Label
Where it can be established that such ineligible features/areas existed in previous years, there is a requirement to reduce the area and apply the relevant financial reduction and/or penalty to the herdowner.	Incentive
EU CAP direct payments provide vital income support for farmers, and act as an important cushion against commodity price volatility.	Non-incentive

Table 5: Sentences from the binary dataset correctly and incorrectly classified with our best-performing model

embedding models and classifier were limited by the small size and class imbalances of the dataset.

We also noted that the pipeline may benefit from span extraction and classification rather than simple sentence classification, especially in cases where there are several incentives contained in a single sentence. We are now experimenting with the POLIANNA dataset for extracting spans to use as features in the sentence classification, enabling multilabel classification for incentives as well (Sewerin et al., 2023).

Ethics

The data for this project consists of sentences from public policy documents, none of which contain private or personal information. Bias may be introduced to the dataset and resulting models by the aforementioned single annotator and annotation decisions in ambiguous cases, as well as by the small size and class imbalance of the dataset. While our pipeline is primarily designed and intended for policy analysis, it could be manipulated to create biased classifiers that mislabel incentives in order to fit a certain agenda, potentially over-representing or under-representing different incentive classes in order to sway the downstream choice of policy solutions. As such, it is important for future implementations of the pipeline to maintain openness and transparency in the construction of their datasets and training of their models.

Acknowledgments

Thank you to Dr. Yifan Wang for annotating a subset of our data for validation. This work was funded by Research Ireland as part of Grant Number SFI/12/RC/2289_P2 Insight_2, the Insight Research Ireland Centre for Data Analytics.

References

Bertrand Badie, Dirk Berg-Schlosser, and Leonardo Morlino. 2011. *International Encyclopedia of Political Science*, volume 1. SAGE Publications, Inc., Thousand Oaks, California.

John Brandt. 2019. Text mining policy: Classifying forest and landscape restoration policy agenda with neural information retrieval. In *KDD Fragile Earth workshop (FEED 2019)*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 2185–2194.

Daniel Firebanks-Quevedo, Jordi Planas, Kathleen Buckingham, Cristina Taylor, David Silva, Galina Naydenova, and René Zamora-Cristales. 2022. Using machine learning to identify incentives in forestry policy: Towards a new paradigm in policy analysis. *Forest Policy and Economics*, 134:102624.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017. Cross-lingual classification of topics in political texts. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 42–46, Vancouver, Canada. Association for Computational Linguistics.

Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.

Steve R. Gunn. 1998. Support vector machines for classification and regression. Project report, University of Southampton. Address: Southampton, U.K.

Loni Hagen, Özlem Uzuner, Christopher Kotfila, Teresa M. Harrison, and Dan Lamanna. 2015. Understanding citizens' direct policy suggestions to the federal government: A natural language processing and topic modeling approach. In *2015 48th Hawaii International Conference on System Sciences*, pages 2134–2143.

Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Ian Magnusson, Noah A. Smith, and Jesse Dodge. 2023. Reproducibility in NLP: What have we learned from the checklist? *Findings of the Association for Computational Linguistics: ACL 2023*, page 12789–12811.

Laura K. Nelson, Derek Burk, Marcel Knudsen, and Leslie McCall. 2021. The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research*, 50(1):202–237.

Jiang Ningpeng, Han Tian, Wang Haibo, Xu Ruzhi, and Ma Shiyu. 2024. A study on structured text parsing for policies based on BERTopic. In *2024 IEEE 6th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, volume 6, pages 16–22.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ludovic Rheault and Christopher Cochrane. 2020. Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, 28(1):112–133.

Sebastian Sewerin, Lynn H. Kaack, Joel Küttel, Fride Sigurdsson, Onerva Martikainen, Alisha Esshaki, and Fabian Hafner. 2023. Towards understanding policy design through text-as-data approaches: The policy design annotations (polianna) dataset. *Scientific Data*, 10(896).

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Zhanna Terechshenko, Fridolin Linder, Vishakh Padmakumar, Michael Liu, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2020. A comparison of methods in political science text classification: Transfer learning language models for politics. *Other Information Systems & eBusiness eJournal*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

John Wilkerson and Andreu Casas. 2017. Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, 20(Volume 20, 2017):529–544.

Shehel Yoosuf and Yin Yang. 2019. Fine-grained propaganda detection with fine-tuned BERT. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 87–91, Hong Kong, China. Association for Computational Linguistics.

Artur Żółkowski, Mateusz Krzyziński, Piotr Wilczyński, Stanisław Giziński, Emilia Wiśnios, Bartosz Pieliński, Julian Sienkiewicz, and Przemysław Biecek. 2022. Climate policy tracker: Pipeline for automated analysis of public climate policies. In *Tackling Climate Change with Machine Learning Workshop*.

A Policy Sources

Table 6 shows the departments of origin for the policies in the dataset at the time of collection. Since the collection however, two of the departments have been renamed and their websites and URLs restructured. In order to address this issue of changing PDF sources and addresses, we provide a ZIP file[9] of the policies in our repository.

---

[9] `https://github.com/mawaskow/policy-classifier/blob/main/policy_scraping/policy_scraping/outputs/forestry/full.zip`

# Policies	Department
45	Agriculture, Food and the Marine
45	Rural and Community Development
44	The Environment, Climate and Communications
4	Housing, Local Government and Heritage

Table 6: The number of policies from each Irish government department represented in the dataset

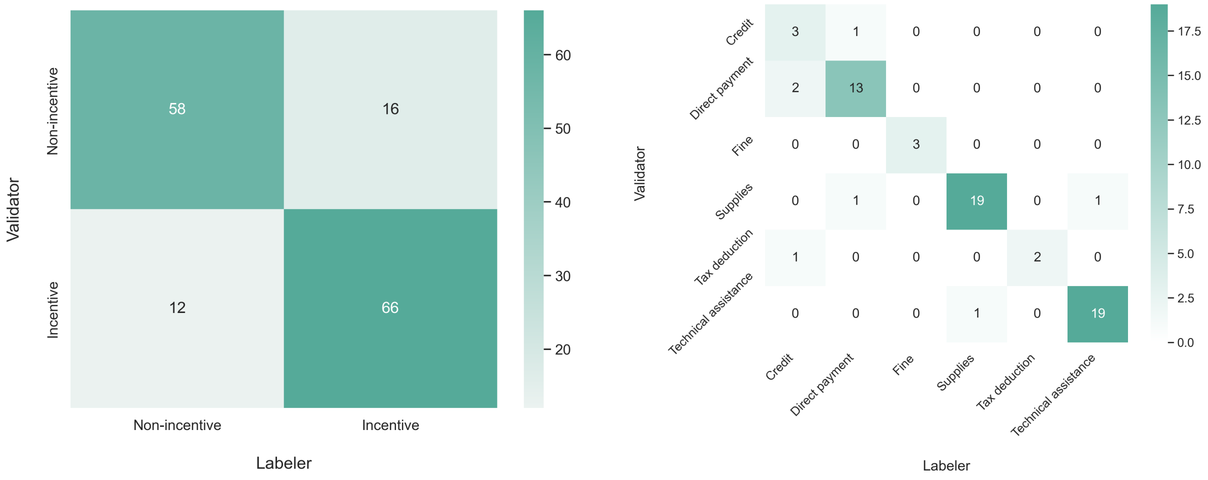


Figure 2: Confusion matrices for the validation sample (left: binary, right: multiclass)

B External Validation

For the external validation of our dataset, we provide the confusion matrices of our subset's labelling in Figure 2 for binary and multiclass classification. Additional information about annotation guidelines can be found on the dataset page.

C Validation Performance

Table 7 reflects performance using the transformer head for classification rather than the SVM classifier used in our main pipeline, resulting in the lower scores observed here. Despite this, the results confirm that MPNet consistently outperforms the other models in incentive sentence embedding for both binary and multiclass classification.

D Precision and Recall

In addition to reporting the overall and label-specific F1 scores of our models' binary and multiclass classification experiments in Tables 2, 3, and 4, we report their precision in Tables 8, 10, and 12, and recall in Tables 9, 11, and 13.

E Ambiguous Cases

Table 14 provides further examples of ambiguous policy sentences.

Model		Binary	Multiclass
XLM-R	d	88.4 ± 2.7	87.2 ± 3.5
	l	88.0 ± 2.6	89.3 ± 3.0
	o	89.1 ± 2.2	91.5 ± 3.7
MPNet	d	**89.8 ± 2.7**	89.0 ± 2.9
	l	89.2 ± 2.2	91.7 ± 3.0
	o	89.1 ± 2.0	**91.8 ± 3.1**
GTE	d	88.4 ± 2.5	73.5 ± 7.3
	l	88.2 ± 2.6	82.9 ± 5.1
	o	89.4 ± 2.3	90.5 ± 2.8
E5	d	88.0 ± 2.8	83.6 ± 3.1
	l	88.1 ± 2.2	88.1 ± 5.0
	o	89.6 ± 2.1	90.4 ± 3.6

Table 7: Overall validation F1 scores of the fine-tuned models (d: default, l: weighted loss, o: oversampling) averaged across 10 different dataset splits

Model		Binary	Multiclass
XLM-R	d	87.7 ± 1.4	94.6 ± 3.8
	l	87.5 ± 1.9	95.0 ± 3.0
	o	87.4 ± 1.4	95.3 ± 3.7
MPNet	d	**88.3 ± 1.6**	96.3 ± 2.1
	l	87.9 ± 2.2	96.6 ± 2.3
	o	88.1 ± 1.0	**97.0 ± 1.9**
GTE	d	87.1 ± 1.1	95.8 ± 3.1
	l	87.2 ± 1.4	95.7 ± 2.6
	o	87.6 ± 1.3	96.7 ± 2.6
E5	d	87.3 ± 1.6	96.0 ± 2.8
	l	86.7 ± 1.7	96.1 ± 1.9
	o	86.3 ± 2.3	95.5 ± 2.9

Table 8: Overall precision scores of the fine-tuned models (d: default, l: weighted loss, o: oversampling) averaged across 10 different dataset splits

Model		Binary	Multiclass
XLM-R	d	87.8 ± 1.5	94.2 ± 4.2
	l	87.6 ± 2.0	94.3 ± 3.5
	o	87.9 ± 1.3	94.7 ± 3.5
MPNet	d	88.5 ± 1.5	95.7 ± 2.5
	l	88.2 ± 2.3	96.2 ± 2.5
	o	**88.7 ± 1.0**	**96.8 ± 2.1**
GTE	d	87.3 ± 1.2	95.7 ± 3.0
	l	87.5 ± 1.4	95.5 ± 2.6
	o	87.9 ± 1.2	96.4 ± 2.6
E5	d	87.6 ± 1.6	95.7 ± 2.9
	l	86.9 ± 1.8	95.8 ± 2.0
	o	86.9 ± 2.1	94.9 ± 3.4

Table 9: Overall recall scores of the fine-tuned models (d: default, l: weighted loss, o: oversampling) averaged across 10 different dataset splits

Model		Non-Incentive	Incentive
XLM-R	d	92.1 ± 0.9	68.5 ± 5.7
	l	92.0 ± 1.1	67.9 ± 6.5
	o	91.0 ± 1.0	71.4 ± 5.0
MPNet	d	**92.3 ± 1.1**	70.6 ± 4.8
	l	92.0 ± 1.4	70.2 ± 7.0
	o	91.4 ± 0.7	**74.0 ± 3.6**
GTE	d	91.6 ± 0.8	67.6 ± 5.7
	l	91.5 ± 1.0	68.2 ± 5.1
	o	91.7 ± 1.1	69.8 ± 4.9
E5	d	91.5 ± 1.1	68.8 ± 5.2
	l	91.3 ± 0.9	66.4 ± 6.1
	o	90.6 ± 1.4	67.7 ± 6.9

Table 10: Label-specific precision scores for the binary dataset (d: default, l: weighted loss, o: oversampling) averaged across 10 different dataset splits

Model		Non-Incentive	Incentive
XLM-R	d	93.0 ± 1.7	65.3 ± 4.2
	l	92.9 ± 1.8	64.7 ± 5.0
	o	94.5 ± 1.4	59.4 ± 4.9
MPNet	d	93.6 ± 1.4	**66.0 ± 5.5**
	l	93.6 ± 2.1	64.5 ± 6.7
	o	**95.1 ± 0.9**	60.8 ± 3.5
GTE	d	92.9 ± 1.8	62.6 ± 4.1
	l	93.2 ± 1.6	62.3 ± 5.1
	o	93.6 ± 1.5	63.0 ± 5.4
E5	d	93.5 ± 1.4	62.3 ± 5.1
	l	92.7 ± 1.7	61.7 ± 4.0
	o	93.6 ± 1.6	57.5 ± 6.8

Table 11: Label-specific recall scores for the binary dataset (d: default, l: weighted loss, o: oversampling) averaged across 10 different dataset splits

Model		Credit	Direct Payment	Fine	Supplies	Tax Deduction	Technical Assistance
XLM-R	d	**100.0 ± 0.0**	93.4 ± 5.0	98.3 ± 5.0	96.1 ± 6.1	78.3 ± 35.0	92.5 ± 5.5
	l	**100.0 ± 0.0**	91.5 ± 7.0	**100.0 ± 0.0**	96.9 ± 4.0	78.3 ± 35.0	93.8 ± 5.4
	o	**100.0 ± 0.0**	**95.9 ± 4.2**	98.3 ± 5.0	96.9 ± 4.1	73.3 ± 35.1	92.2 ± 6.4
MPNet	d	**100.0 ± 0.0**	91.9 ± 6.5	**100.0 ± 0.0**	**98.2 ± 2.8**	93.3 ± 20.0	95.9 ± 5.7
	l	97.5 ± 7.5	91.9 ± 6.7	**100.0 ± 0.0**	**98.2 ± 2.8**	95.0 ± 15.0	97.4 ± 3.3
	o	97.5 ± 7.5	95.4 ± 5.0	**100.0 ± 0.0**	97.0 ± 3.0	95.0 ± 15.0	**97.5 ± 3.1**
GTE	d	98.0 ± 6.0	91.2 ± 7.3	**100.0 ± 0.0**	97.6 ± 3.8	90.0 ± 30.0	96.1 ± 4.3
	l	98.0 ± 6.0	92.7 ± 6.9	**100.0 ± 0.0**	97.6 ± 3.8	90.0 ± 30.0	94.5 ± 5.6
	o	**100.0 ± 0.0**	95.9 ± 4.1	**100.0 ± 0.0**	96.9 ± 3.1	83.3 ± 34.2	95.8 ± 4.6
E5	d	98.0 ± 6.0	95.1 ± 5.5	**100.0 ± 0.0**	97.6 ± 3.9	95.5 ± 15.0	93.1 ± 5.8
	l	98.0 ± 6.0	94.5 ± 5.0	**100.0 ± 0.0**	97.0 ± 4.0	**100.0 ± 0.0**	94.3 ± 4.3
	o	**100.0 ± 0.0**	94.5 ± 6.1	**100.0 ± 0.0**	97.5 ± 3.0	93.3 ± 20.0	91.7 ± 6.8

Table 12: Label-specific precision scores for the multiclass dataset (d: default, l: weighted loss, o: oversampling) averaged across 10 different dataset splits

Model		Credit	Direct Payment	Fine	Supplies	Tax Deduction	Technical Assistance
XLM-R	d	95.0 ± 10.0	93.3 ± 6.2	90.0 ± 13.4	95.6 ± 4.9	90.0 ± 30.0	94.7 ± 7.2
	l	95.0 ± 10.0	93.3 ± 9.0	86.0 ± 15.6	96.3 ± 5.0	90.0 ± 30.0	96.0 ± 5.3
	o	**97.5 ± 7.5**	93.3 ± 9.0	92.0 ± 13.3	95.6 ± 4.9	90.0 ± 30.0	95.3 ± 5.2
MPNet	d	90.0 ± 12.2	95.8 ± 6.7	96.0 ± 12.0	96.3 ± 5.0	**100.0 ± 0.0**	96.0 ± 5.3
	l	90.0 ± 12.2	**96.7 ± 4.1**	96.0 ± 8.0	**98.8 ± 2.5**	**100.0 ± 0.0**	94.7 ± 7.2
	o	92.5 ± 11.5	95.8 ± 4.2	98.0 ± 6.0	98.1 ± 2.9	**100.0 ± 0.0**	96.7 ± 5.4
GTE	d	92.5 ± 11.5	**96.7 ± 4.1**	**100.0 ± 0.0**	95.6 ± 2.9	90.0 ± 30.0	94.7 ± 5.0
	l	85.0 ± 16.6	**96.7 ± 4.1**	**100.0 ± 0.0**	96.9 ± 3.1	90.0 ± 30.0	94.7 ± 5.8
	o	95.0 ± 10.0	95.0 ± 6.7	96.0 ± 12.0	97.5 ± 3.1	90.0 ± 30.0	**97.3 ± 4.4**
E5	d	92.5 ± 11.5	94.2 ± 6.5	96.0 ± 8.0	96.9 ± 3.1	**100.0 ± 0.0**	96.0 ± 4.4
	l	90.0 ± 12.2	95.0 ± 4.1	**100.0 ± 0.0**	96.9 ± 4.2	**100.0 ± 0.0**	95.3 ± 4.3
	o	92.5 ± 11.5	92.5 ± 6.9	96.0 ± 12.0	95.6 ± 4.9	**100.0 ± 0.0**	96.0 ± 5.3

Table 13: Label-specific recall scores for the multiclass dataset (d: default, l: weighted loss, o: oversampling) averaged across 10 different dataset splits

The Eco-Schemes provides supports to farmers who undertake specific agricultural practices, including extensive farming, tree planting, sewing a multi-species sward, and enhancing crop diversification.

Department of Finance and DAFM to improve dissemination of information on taxation incentives including engagement with advisors, tax consultants and accountants.

LEADER may provide support rates greater than 65% in accordance with Article 73(4) (c)(ii) where investments include basic services in rural areas and infrastructure in agriculture and forestry, as determined by Member States.

Table 14: Ambiguous policy sentences, unclear in either their status as incentive or non-incentive, or in the type of incentive they are.

Narrative Structure Extraction in Disinformation and Trustworthy News: A Comparison of LLM, KG, and KG-Augmented Pipelines

Justina Mandravickaitė
Vytautas Magnus University, Lithuania
`justina.mandravickaite@vdu.lt`

Abstract

Identifying narrative structures such as characters, events, causes, and frames in political news is essential to detecting bias and understanding political dynamics, among other areas. Large language models (LLMs), while performing well on a variety of natural language processing (NLP) tasks, may hallucinate, whereas pure knowledge graph (KG) methods, while excelling at text structuring and information extraction, suffer from sparsity. In this study, three pipelines for extracting narrative structures from disinformation and trustworthy news are evaluated: (1) LLM-only prompt-based extraction, (2) KG-only graph-based queries and (3) an augmented approach combining LLM prompts augmented with appropriate KG data. The results were evaluated intrinsically as well as extrinsically. For intrinsic evaluation, element coverage, fuzzy overlap, coherence, coverage gain and error reduction were measured, while extrinsic evaluation used matching with Wikidata and downstream classification. The augmented pipeline improved the coverage and coherence of narrative elements, but also boosted the classification of disinformation, as it outperformed both LLM-only and KG-only pipelines.

1 Introduction

Narrative extraction from political news is important for revealing how disinformation manipulates the stories by, e.g., assigning blame (Rauchfleisch and Jungherr, 2024), inflating or downplaying events (Keith Norambuena et al., 2023) or constructing false causal chains (Lei and Huang, 2023) in order to shape public perception. Identification of narrative structures aids in revealing rhetorical strategies that fact-checking alone may miss (Byram, 2022).

Large language models (LLMs), while perform well in diverse NLP tasks, including information extraction, may hallucinate relations or entities (Mckenna et al., 2023; Li et al., 2024a). Meanwhile, knowledge graphs (KGs) can suffer from incomplete coverage (Wang et al., 2022) or outdated ontologies (Zhou et al., 2024; Hegde et al., 2025), among other issues, which may result in, e.g., missing events (Zhuang et al., 2023) or misclassified roles (Lu et al., 2024). While many pipelines are assessed using intrinsic metrics (e.g., coherence (German et al., 2025) or reconstruction accuracy (Keith Norambuena et al., 2023)) or individual extrinsic tasks (e.g., downstream classification (Das et al., 2024)), there is an opportunity to develop more comprehensive validation protocols that combine both approaches to provide a more comprehensive evaluation.

To address these issues, in this study, three pipelines for extracting narrative structures from disinformation and trustworthy news are evaluated: (1) LLM-only prompt-based extraction, (2) KG-only graph-based queries and (3) an augmented approach combining LLM prompts augmented with appropriate KG data. The proposed evaluation framework combines intrinsic and extrinsic evaluation. It integrates measures of narrative coverage, fuzzy overlap, coherence, coverage gain and error reduction (intrinsic), alongside downstream tasks such as matching with Wikidata and disinformation classification (extrinsic).

2 Related Work

Advances in narrative extraction for political and social science have relied on structured representations such as event schemas to analyze social processes, events and framing (Zhang et al., 2019; Halterman, 2020; Li et al., 2024b). Transformer-based multilingual models have achieved high accuracy in crisis event detection across languages, illustrating usefulness of schema-guided extraction (Hettiarachchi et al., 2021).

Schema refinement has emerged for extracting

Proceedings of the 5th Workshop on Computational Linguistics for the Political and Social Sciences (CPSS), pages 80–97
September 10-11, 2025 ©2025 Association for Computational Linguistics

events without predefined templates, such as Liberal Event Extraction (LEE) which jointly extracts events and induces schemas (Li and Geng, 2024). Integrating schemas with cultural norms and commonsense knowledge was used to support framing analysis (Li et al., 2024b). Visualization tools have been used to reveal how discrete events affect outcomes over time (Shen et al., 2024).

Also, graph-based methods have been proposed, such as using AMR (abstract meaning representation) to extract actors, events, and perspectives from digital media(Pournaki and Willaert, 2024). Knowledge Graph–based approaches enhance semantic precision and reveal causal relationships in narrative analysis (de Kok et al., 2024; Romanou et al., 2023). In bias detection, integrating frame-based knowledge with text models improved bias and stance detection (Li, 2021). Also, domain-specific KGs for news facilitate advanced bias detection and narrative synthesis (Yang et al., 2024).

Event-centric narrative extraction has been a trend for a while (Voskarides et al., 2021; Keith Norambuena et al., 2023). Advances and challenges in multimodal event extraction has been introduced as well (Hürriyetoğlu et al., 2024).

Recent research revealed that hybrid methods mitigate weaknesses of purely symbolic (such as KG-based) or neural approaches (such as LLM-based) (Panchendrarajan and Zubiaga, 2024; Zhu, 2024). For example, multi-agent approaches has been recommended for KG construction and reasoning, (Zhu et al., 2024). Also, at least several LLM–KG integration frameworks have been outlined, such as SymAgent, a neural-symbolic agent for multi-step reasoning and automatic KG updates (Liu et al., 2025)or MindMap which merges KGs and LLMs for improved inference transparency (Wen et al., 2024).

Different proposed directions also involve merging LLMs with relation extraction to build semantically rich KGs (Santini, 2024) and optimize LLM reasoning over KGs via selective triple selection (Wang, 2025). Also, recent developments in processing long narratives address past limitations by integrating dynamic KGs with LLMs to improve story comprehension (Andrus et al., 2022).

Narrative extraction evaluation includes variety of metrics and approaches,such as intrinsic metrics for event and attribution extraction, stressing reproducible practices (Zhang et al., 2019). Also, the complexity of assessing narrative coherence has been highlighted (Santana et al., 2023). Furthermore, studies have evaluated narrative elements via Multiple Choice Narrative Cloze (Hatzel and Biemann, 2023), accuracy metrics and downstream performance (Tang et al., 2021), structural coherence checks and user studies (Norambuena and Mitra, 2021).

This paper reports comparative evaluation of LLM-only, KG-only and Augmented pipelines for extracting narrative structures in political news, assessed with both intrinsic metrics and extrinsic grounding. Unlike typical KG–LLM work, in this paper narrative extraction gains were directly linked to improved disinformation classification while maintaining interpretable representations.

3 Methods

3.1 Data

For this study a part of the dataset for multilingual detection of pro-Kremlin disinformation in news articles (Leite et al., 2024), containing data on disinformation and trustworthy news articles, was used. This dataset consists of 18 249 articles in 42 languages published January 2015-July 2023. As the full text of the news articles was not publicly available, to reconstruct the dataset URL links of the articles were used with DiffBot API[1] (free for academic purposes) to acquire them. Only English-language articles were selected, i.e. a total of 6,546 articles (425 – disinformation and 6121 – trustworthy). Some articles were no longer available or have been modified. So, after filtering and cleaning the final dataset used in this study was made of 308 disinformation news articles and 302 – trustworthy news articles. Labeling news articles as 'disinformation' and 'trustworthy' is based on the original dataset, i.e., article labels were reused as they were assigned in the original dataset (Leite et al., 2024). More details of the final dataset are presented in Appendix B.

While the data is of modest size, the experimentation targets pipeline comparison under controlled conditions, not training a large downstream model. Although the final dataset is $\sim$ 300 articles per class, intrinsic evaluations operate per article and summarize many extracted elements. Also, extrinsic evaluation analyses use simple, regularized classifier (logistic-regression) with grouped, repeated cross-validation to reduce variance.

---

[1] Accessible at `https://www.diffbot.com/`

3.2 Pipelines

Based on existing narratology and framing research, the following narratives structures were extracted and compared in three pipelines (LLM-only, KG-only and Augmented): characters (entities – persons and organizations), events (predicates indicating what happens to/among characters over time), causal links (predicates that express cause/effect), framing (predicates that indicate attribution/association) (Hellman, 2024; Heddaya et al., 2024; Wang et al., 2025).

For this study, the first step, before applying any of the pipelines, was text summarization. This step was applied to reduce noise and to use computational resources more efficiently in the later stages. Extractive summarization was used for this task as it extracts the most important sentences from the given text, preserving factuality and original wording (Hofmann-Coyle et al., 2022). Article summaries were used in all three pipelines that were comprised of different steps.

Summarization. The cleaned news articles were summarized using two extractive summarization approaches – LexRank Summarizer (Erkan and Radev, 2004) and BERT Extractive Summarizer (Miller, 2019). To choose from two summary-candidates, mean ROUGE-1/2/L F1 scores (Lin, 2004) were computed against the original article text. ROUGE-1/2/L F1 score favors coverage of source content and discourages off-source material and was used as a heuristic to choose between two summary-candidates. Therefore, summaries generated by the approach that had higher ROUGE values, showing higher source-overlap, were used for the experiments.

The length of the extractive summary was tied to text length: texts under 300 words were summarized in two sentences; texts of 300–799 words long – in four sentences and texts of 800 words or longer – in six sentences. To ensure quality, summaries were also manually inspected, correcting such issues as occasional incomplete final sentences. These summaries were used with all three narrative extractive pipelines. The examples of the summaries are presented in Appendix A.

LLM-only pipeline. Characters, events, causal links and framing were extracted via prompting. For this task, Mistral Small 3.1 24b[2] was used. The decision was made to use a smaller model, not the largest and most popular one like OpenAI's ChatGPT to manage computational and financial costs while maintaining competitive performance. The prompt used for extraction is presented in Appendix C.

KG-only pipeline. This pipeline consisted of several components, introduced below.

1. *Relational triple extraction.* The pre-trained REBEL (Relation Extraction By End-to-end Language generation) (Cabot and Navigli, 2021) model was applied for this task, extracting entity-relation-entity triples from summaries of news articles. These triples were used for graph construction. REBEL combines Named Entity Recognition (NER) and Relation Classification (RC) into a single task and covers >200 relationship types. Extracted triples were aggregated per article and manually validated.

2. *Mapping and inferring relations.* Relations of REBEL schema that imply framing, events and causality were mapped and inferred manually (for details, see Appendix D). This enrichment was necessary to make outputs of all there pipelines comparable.

3. *Building RDF graphs.* In this step, one rdflib graph per article (i.e. from only that article's triples) was built. In this graph structure entities (nodes) of the triples were linked by their relations (edges). Python library rdflib[3] was used for this task.

4. *Querying narrative elements.* REBEL triples (entity-relation-entity) were grouped into four relation families that were used to compute structure-level features: PERSON_RELS (relations that assert roles, affiliations or actor–actor ties), EVENT_RELS (relations that indicate actions/occurrences or changes; also used to collect event nodes (instance_of = Event)), CAUSAL_RELS (explicit cause/effect predicates and a small set of implied causal links as a result from mapping and inferring relations in previous step), FRAMING_RELS) (attribution/stance and portrayal; REBEL labels were mapped to short paraphrases (e.g., described_as → "is portrayed as"), see Appendix D). These narra-

---

[2] Accessible at https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503

[3] Accessible at https://github.com/RDFLib/rdflib

tive element types were extracted via querying over RDF graphs.

Augmented pipeline. For extraction of narrative elements (characters, events, causal links and framing) from article summaries, prompts were augmented with relevant data from article-level KGs to provide context. The prompt used for extraction of narrative elements with Augmented pipeline is presented in Appendix C.

3.3 Evaluation

The evaluation framework of this study combines intrinsic as well as extrinsic evaluation to get a more comprehensive assessment of extraction of narrative structures. The special focus in this evaluation is paid to exploring the use of an LLM and KG in combination. Also, formal paired significance tests were applied only for metrics that are directly comparable across pipelines across pipelines and quantify extraction quality rather than volume, i.e., overlap (fuzzy Jaccard), coverage gain and error reduction, and extrinsic F1 in the classification. For other quantities, descriptive statistics are reported, avoiding inferential claims.

3.3.1 Intrinsic Evaluation

Intrinsic evaluation of outputs of all three pipelines consisted of the following components:

Coverage: raw counts as well as average counts per element were calculated. This measures recall potential by element type and is simple, interpretable, surfaces systematic under-extraction.

Fuzzy-Jaccard overlap: paraphrase-aware intersection / union between pipelines in terms of extracted narrative elements was calculated (Cross et al., 2020). For this task, *multi-qa-MiniLM-L6-cos-v1*[4] model and cosine similarity (Gunawan et al., 2018) with threshold 0.7 were used. Fuzzy Jaccard overlap shows cross-pipeline agreement on which elements were extracted, while tolerating paraphrases. This measure is scale-free, directly comparable and allows for semantic matching that avoids penalizing lexical variation.

Mean coherence: embedding-based sentence similarity was calculated for LLM-only and Augmented pipelines. Items of the lists of extracted narrative elements were treated as standalone "sentences." For this task, *all-MiniLM-L6-v2*[5] model was used and average adjacent-sentence cosine similarity was computed between items in the lists of narrative elements extracted from the same summaries of the news articles. For KG-only pipeline, graph-based (structural) coherence metrics (density (Shang and Huang, 2024), avg. path length and largest component ratio) were calculated for the graphs, created from relational triples, belonging to each article in the dataset. Semantic (embedding-based) coherence may show internal thematic consistency of extracted elements. Meanwhile, structural (graph-based) coherence may reveal narrative connectedness or fragmentation in the triple graph.

Orthogonal metrics (coverage gain & error reduction): these metrics were used for LLM-only and Augmented pipelines to compare and evaluate coverage of narrative elements and find unmatched elements, which may indicate model hallucination. These metrics were calculated using the following formulas:

Coverage gain for an element type = | Augmented pipeline | – | LLM-only pipeline |

Error reduction = | LLM-only pipeline unmatched elements | – | Augmented pipeline unmatched elements

Orthogonal metrics were calculated per element type. Coverage gain is the corpus-level difference in total extracted items. Unmatched elements are items produced by a pipeline that have no fuzzy match in the other pipeline. Error reduction refers to the decrease in unmatched elements when switching from LLM-only to Augmented pipeline. A ratio > 1 of Coverage gain and Error reduction means that Augmented pipeline adds more new correct elements than LLM-only pipeline removes. Meanwhile, a ratio < 1 means Augmented pipeline introduces more "errors" (unmatched elements) than additional coverage, indicating added noise. Orthogonal metrics capture the augmentation trade-off between newly matched and unmatched elements. In other words, these metrics directly quantify, per element type, whether augmentation was justified.

3.3.2 Extrinsic Evaluation

Loose matching with Wikidata: For a grounding / checking consistency, narrative elements, extracted with all three pipelines, were matched with loosely Wikidata. Linking extracted elements known knowledge base tests whether outputs correspond to real entities, reducing hallucinations and improving interpretability and Wikidata is widely

---

[4] Accessible at `https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-cos-v1`

[5] Accessible at `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

used for such tasks (Möller et al., 2022). A stratified sample of 300 elements from each pipeline was taken. The unique labels were used to query against Wikidata[6] to determine if they matched real-world entities. For this task, Wikidata Search API (based on the MediaWiki API)[7] was used to perform a relevance-based search (or loose matching), i.e., matches were based on labels, aliases, redirects, and included some tolerance for minor typo. The results were ranked by relevance.

Downstream classification: Logistic-regression classifier was built for this task (see Appendix E for details). Counts of narrative elements and coherence (for LLM-only and Augmented pipelines – embedding-based metric and for KG-only pipeline – graph-based metrics) were used as features. For evaluation, 5-fold and 10-fold cross-validation(Wong and Yeh, 2019) was performed and F1-score (Naidu et al., 2023) was used to assess the results. Classification was used as an extrinsic evaluation to assess which narrative extraction pipeline more effectively captures the signal needed to distinguish disinformation from trustworthy news.

Paired tests (paired t-test (Rainio et al., 2024) and Wilcoxon signed-rank test (Rey and Neuhäuser, 2011)) were used to evaluate the statistical significance of the results by performing pairwise comparisons on the pipelines.

Cohen's d (Goulet-Pelletier and Cousineau, 2018) was calculated to measure the effect size, i.e. how big is the difference in classification performance by LLM-only, KG-only and Augmented pipelines when distinguishing disinformation from trustworthy news.

4 Results

4.1 Intrinsic Quality

4.1.1 Raw Counts and Mean Coverage of Narrative Elements

Table 1 shows descriptive counts of extracted narrative elements per pipeline. These raw counts summarize extraction tendencies and are not subjected to formal significance testing. As Table 1 shows, KG-only pipeline lags behind on every element, reflecting that raw triple extraction alone captures only a small part of narrative content.

Meanwhile, Augmented pipeline slightly outperformed LLM-only pipeline on characters and events for both classes. As for causal links, slightly higher counts resulted from Augmented pipeline for disinformation. However, for trustworthy news LLM-only pipeline provided higher counts, though the difference, in comparison to Augmented pipeline, is very small. Meanwhile, for framing more elements were extracted with LLM-only pipeline for both disinformation and trustworthy news.

Trustworthy articles generally provided higher counts and means than disinformation across all pipelines. This could be likely because longer, more structured reporting offers richer narratives.

The augmented pipeline mostly maintained LLM-only breadth of coverage while adding a modest boost, such as on actors and events, over LLM-only. KG alone was too sparse for standalone narrative extraction.

4.1.2 Overlap of Narrative Elements by Pipeline

To assess overlap of extracted narrative elements between pipeline pairs on the same article and element type per class, fuzzy Jaccard was used. As presented in Table 2, there almost no overlap for **llm∩kg**, meaning each alone extracts mostly disjoint sets of elements. Meanwhile, **llm∩augmented** overlap more substantial (0.46–0.47 for characters, 0.13–0.18 for events, up to 0.15 for framing). This shows that Augmented pipeline not only retained most LLM-only elements but also enriched them with additional ones from KG-only.

However, **kg∩augmented** overlap was low, though not so low as for **llm∩kg**, which means that Augmented pipeline's additional KG-based data was a small part of its output.

4.1.3 Coherence Assessment

To assess semantic continuity (for LLM-only and Augmented pipelines) and structural cohesion (for KG-only pipeline) within output for the same article, mean coherence measures were calculated. Embedding-based (semantic) coherence was slightly higher in Augmented pipeline than LLM-only one for disinformation (0.411 vs. 0.400) and there was no difference in coherence in terms of trustworthy news (Table 3).

Graph-based (structural) coherence in the KG-only pipeline showed that, in disinformation arti-

---

[6] Accessible at `https://www.wikidata.org/wiki/Wikidata:Main_Page`

[7] Accessible at `https://www.wikidata.org/w/api.php?action=help`

Element type	**Pipeline**	**Disinformation**	**Trustworthy**
Characters / actors	LLM-only	2226 (7.25)	2453 (8.10)
	KG-only	1147 (3.74)	1228 (4.05)
	Augmented	**2377 (7.74**	**2541 (8.39)**
Events	LLM-only	1826 (5.95)	2043 (6.74)
	KG-only	244 (0.79)	291 (0.96)
	Augmented	**1849 (6.02)**	**2118 (6.99)**
Causal links	LLM-only	1287 (4.19)	**1362 (4.50)**
	KG-only	71 (0.23)	77 (0.25)
	Augmented	**1298 (4.23)**	1355 (4.47)
Framing (actors+events)	LLM-only	**2401 (7.82)**	**2462 (8.13)**
	KG-only	106 (0.35)	158 (0.52)
	Augmented	2227 (7.25)	2274 (7.50)

Table 1: Raw counts and means (in brackets)

Element	**llm∩kg (D/T)**	**llm∩augmented (D/T)**	**kg∩augmented (D/T)**
Characters/actors	0.14 / 0.13	**0.47 / 0.46**	0.19 / 0.19
Events	0.008 / 0.006	**0.18 / 0.13**	0.013 / 0.014
Causal links	0 / 0	**0.076 / 0.058**	0.002 / 0.002
Framing	0.002 / 0.001	**0.15 / 0.12**	0.001 / 0.003

Table 2: Mean Fuzzy Jaccard (D – disinformation, T – trustworthy news)

cles, nearly 43 % of nodes belonged to the largest connected component, compared to 38 % in trustworthy articles. The average shortest path length was approximately 1.2 in both cases, and graph density was higher for disinformation (0.23 vs. 0.20). These results suggested that narrative elements extracted from disinformation articles were more interconnected. The similar average path lengths indicate that, despite differences in overall connectedness, the "distance between ideas" was similar across both categories.

The augmented pipeline's modestly higher coherence suggested that incorporating KG-derived facts helped to make the narrative elements semantically tighter, even if the gain was small. In contrast, KG-only graphs presented measurable structural coherence, though their metrics were not directly comparable to embedding-based measures.

4.1.4 Coverage Gain and Error Reduction

Orthogonal metrics (coverage gain and error reduction) were computed only for the LLM-only and Augmented pipelines. The KG-only pipeline was excluded because its extraction process was based on querying structured relational graphs which is very different from the prompt-based extraction in the LLM-only and Augmented pipelines.

In terms for disinformation, for **actors/characters** and **events**, the Augmented pipeline gives a modest but consistent net boost in coverage (2–7 %) with minimal noise (Table 4).

For **causal links**, the Augmented pipeline captured just as many or very slightly more matched causal links as it introduces unmatched ones.

For **framing edges**, the KG-augmented prompts added quantity but at the cost of introducing more unmatched framing elements, suggesting the prompting strategy for framing relations may need further refinement (e.g. more precise examples or stricter filtering).

Regarding trustworthy news, Augmented pipeline also provided slightly more new **actors/characters** and **events** than mismatches removed (Table 5).

For causal links, augmentation slightly favored mismatch reduction over added coverage. Finally, for framing edges, augmentation primarily improved precision in reducing mismatches than in-

Pipeline	**Metric**	**Disinformation**	**Trustworthy**
LLM-only	Embedding-based coherence	0.400	0.396
Augmented	Embedding-based coherence	**0.411**	**0.396**
KG-only	Largest component ratio	0.427	0.382
	Avg. path length	1.222	1.244
	Density	0.228	0.195

Table 3: Mean coherence by pipeline

Element	Coverage Gain	Error Reduction	Gain / Reduction
Characters/actors	1151	1079	1.07
Events	1736	1703	1.02
Causal links	1233	1226	1.01
Framing edges	2291	2460	0.93

Table 4: LLM-only vs. Augmented pipeline coverage gain & error reduction: disinformation news

Element	Coverage Gain	Error Reduction	Gain / Reduction
Characters/actors	1289	1252	1.03
Events	1996	1933	1.03
Causal links	1277	1297	0.98
Framing edges	2306	2567	0.90

Table 5: LLM-only vs. Augmented pipeline coverage gain & error reduction: trustworthy news

creasing in new elements. This, again, suggests that the prompting strategy for framing relations may need further tuning.

Overall, these intrinsic evaluations supported the assumption that Augmented pipeline outperforms pure LLM-based extraction on core narrative elements, especially actors and events. Meanwhile, framing may need a more targeted prompt design or post-extraction filtering.

4.2 Extrinsic Quality

4.2.1 Matching with Wikidata

KG-only extractions of narrative elements aligned with Wikidata at a very high rate ($\sim$ 70 %), reflecting that REBEL triples largely take out canonical entities and relations that already exist in Wikidata (Table 6). Meanwhile, LLM-only was far less grounded ($\sim$ 15–20 %), since free-form prompts often generate paraphrases, alternate names, or relations that are not directly findable in Wikidata. Finally, Augmented pipeline stayed in between the aforementioned pipelines, reflecting LLM's grounding by including the KG triples in the output.

As for matching per element type, Augmented pipeline outperformed LLM-only in terms of characters/actors but still trailed behind KG-only as $\sim 60\%$ of sampled characters/actors extracted by this pipeline matched Wikidata entities (Table 7).

However, all pipelines struggled in grounding events in Wikidata, though again Augmented pipeline had a higher match rate than LLM-only and KG-only had more matches than both other pipelines.

Finally, causal links and framing had very low Wikidata coverage overall ($\leq 3\%$), reflecting that these relations are rarely modeled as explicit triples in Wikidata. KG-only did not cover these elements at all (Table 7).

4.2.2 Downstream Classification

Document classification was used as an extrinsic evaluation to examine whether pipeline-derived narrative features, rather than surface text, may help distinguish between disinformation and trustworthy news.

Augmented pipeline achieved the highest mean F1-score in both 5- and 10-fold cross-validation, 0.666 and 0.657, accordingly (Table 8). The improvement is moderate in effect size (d $\sim$ 0.6–0.8)

Pipeline	Trustworthy	Disinformation
LLM-only	44 / 300 (14.7 %)	61 / 300 (20.3 %)
Augmented	71 / 300 (23.7 %)	77 / 300 (25.7 %)
KG-only	211 / 300 (70.3 %)	210 / 300 (70.0 %)

Table 6: Overall Wikidata matches

Element type	Pipeline	Trustworthy (out of sample)	Disinfo (out of sample)
Characters/actors	LLM-only	31 / 300 (10.3 %)	47 / 300 (15.7 %)
	Augmented	60 / 300 (20.0 %)	51 / 300 (17.0 %)
	KG-only	178 / 300 (59.3 %)	176 / 300 (58.7 %)
Events	LLM-only	4 / 300 (1.3 %)	6 / 300 (2.0 %)
	Augmented	6 / 300 (2.0 %)	14 / 300 (4.7 %)
	KG-only	33 / 300 (11.0 %)	34 / 300 (11.3 %)
Causal links	LLM-only	7 / 300 (2.3 %)	4 / 300 (1.3 %)
	Augmented	5 / 300 (1.7 %)	9 / 300 (3.0 %)
	KG-only	—	—
Framing	LLM-only	—	4 / 300 (1.3 %)
	Augmented	—	3 / 300 (1.0 %)
	KG-only	—	—

Table 7: Wikidata maches by element type

(Table 11 and shows a clear trend toward statistical significance in the more stable 10-fold cross-validation ($p \approx 0.08$) as showed in Table 10.

Furthermore, Augmented pipeline significantly outperformed KG-only in both cross-validations (t-test $p < 0.05$ for both splits; Wilcoxon significant in 10-fold cross-validation), as presented in Tables 9 and 10, and large effect sizes ($d \sim 0.9$–1.9) as showed in Table 11.

Finally, LLM-only vs. KG-only did not show significant statistical difference in either 5- or 10-fold cross-validation ($p > 0.05$ for all tests) and the effect size was small.

5 Conclusions

The Augmented pipeline showed intrinsic gains in narrative-element coverage and coherence, which also were translated to meaningful downstream improvements in disinformation classification as Augmented pipeline outperformed LLM-only pipeline by a moderate margin and KG-only pipeline – by a large margin, measured in effect size.

Augmented pipeline matched or slightly exceeded LLM-only on characters, events, and causal links while KG-only lagged behind. Also, Augmented pipeline retained most all LLM-only extractions for characters/actors and added KG-informed ones, whereas LLM-only and KG-only barely overlapped.

Moreover, Augmented pipeline's embedding-based coherence was marginally higher than LLM-only, suggesting KG-based augmentation tightened the narrative. Furthermore, Augmented pipeline gained a boost in valid characters and events beyond LLM-only, with minimal noise. Causal links, on the other hand, were similar in quantity, while framing was in need of refinement.

In addition, Augmented pipeline had higher match rate to Wikidata than LLM-only, showing a balance between creative extraction and factual grounding, though lagging behind KG-only due to its schema closely matching Wikidata.

Finally, Augmented pipeline outperformed the other two pipelines in downstream classification – achieved 0.666 ± 0.022 (5-fold cross-validation) and 0.657 ± 0.032 (10-fold cross-validation), versus 0.619/0.607 for LLM-only and 0.591/0.595 for KG-only. In statistical tests Augmented pipeline significantly outperformed KG-only pipeline ($p < 0.05$) and showed a moderate effect size over LLM-only pipeline (Cohen's $d \sim 0.6$–0.8), with a trend toward significance in 10-fold cross-validation ($p \approx 0.08$).

Future work includes experimentation with both smaller- and larger-scale open-source and propri-

Validation	LLM-only	KG-only	Augmented
5-fold	0.619 ± 0.060	0.591 ± 0.025	**0.666 ± 0.02**2
10-fold	0.607 ± 0.082	0.595 ± 0.061	**0.657 ± 0.032**

Table 8: Mean F1-scores)

Pipeline pairs	**Paired t-test**	**Wilcoxon signed-rank**
LLM-only vs KG-only	t = 1.358, p = 0.246	W = 2.000, p = 0.188
LLM-only vs Augmented	t = -1.708, p = 0.163	W = 3.000, p = 0.312
KG-only vs Augmented	t = -4.259, p = 0.013	W = 0.000, p = 0.062

Table 9: Significance testing (5-fold validation)

Pipeline pairs	**Paired t-test**	**Wilcoxon signed-rank**
LLM-only vs KG-only	t = 0.330, p = 0.749	W = 25.000, p = 0.846
LLM-only vs Augmented	t = -1.986, p = 0.078	W = 10.0, p = 0.084
KG-only vs Augmented	t = -2.799, p = 0.021	W = 3.0, p = 0.020

Table 10: Significance testing (10-fold validation)

Comparison	**5-fold cross-validation**	**10-fold cross-validation**
LLM-only vs. KG-only	+0.61 (moderate)	+0.10 (small)
LLM-only vs. Augmented	–0.76 (moderate)	–0.63 (moderate)
KG-only vs. Augmented	–1.90 (**large**)	–0.89 (**large**)

Table 11: Effect Sizes (Cohen's d)

etary models, compare multiple KG-building methods (e.g. pipeline vs. joint extraction) and prompt-engineering strategies as well. The extension to non-English sources by leveraging cross-lingual embeddings and integrating language-specific KGs is also planned. To better capture narrative nuances, human-in-the-loop coherence judgments, causal-structure metrics, fact-checking efficiency, etc. are among future plans as well.

Limitations

Despite encouraging results, this study has several limitations. Narrative quality was evaluated through proxy metrics (coverage, overlap, coherence), but lacked manually annotated ground truth, preventing a direct assessment of factual correctness, hallucination, or omission.

Also, all news articles were in English and focused on disinformation and trustworthy reporting, therefore, results may not generalize to other political genres or languages with different discourse structures.

Furthermore, single LLM (*Mistral Small 3.1 24b*) was used for narrative extraction. Additional experiments are needed to explore whether the observed benefits of Augmented pipeline persist across other LLMs.

Moreover, the downstream task used logistic regression with hand-crafted narrative features. While this helped with interpretability, more complex classifiers might capture additional signals.

In addition, fuzzy Jaccard overlap metric relies on sentence embeddings and thresholds. While it captures surface variation, it may under- or overestimate semantic similarity, especially in terms of figurative or culturally specific language.

Finally, The Augmented pipeline relied on KG-derived context included into the prompt. However, different prompt writing techniques and order of elements making this pipeline (e.g., using LLM outputs to filter KG content) were not tested and the results may be different.

Acknowledgments

This research was funded by the Research Council of Lithuania (LMTLT), grant agreement No. S-PD-24-88.

References

Berkeley R Andrus, Yeganeh Nasiri, Shilong Cui, Benjamin Cullen, and Nancy Fulda. 2022. Enhanced story comprehension for large language models through dynamic document-based knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10436–10444.

Katra Byram. 2022. Narrative as social action: making rhetorical narrative theory accountable to context. *Poetics Today*, 43:455–478.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381.

Valerie Cross, Valeria Mokrenko, Keeley Crockett, and Naeemeh Adel. 2020. Using fuzzy set similarity in sentence similarity measures. In *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–8. IEEE.

Rohan Das, Aditya Chandra, I-Ta Lee, and María Leonor Pacheco. 2024. Media framing through the lens of event-centric narratives. In *Proceedings of the The 6th Workshop on Narrative Understanding*, pages 85–98.

Mike de Kok, Youssra Rebboud, Pasquale Lisena, Raphael Troncy, and Ilaria Tiddi. 2024. From nodes to narratives: A knowledge graph-based storytelling approach. In *TEXT2STORY 2024, 7th International Workshop on Narrative Extraction from Texts (Text2Story), colocated with ECIR 2024*.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Fausto German, Brian Keith, and Chris North. 2025. Narrative trails: A method for coherent storyline extraction via maximum capacity path optimization. *arXiv preprint arXiv:2503.15681*.

Jean-Christophe Goulet-Pelletier and Denis Cousineau. 2018. A review of effect sizes and their confidence intervals, part i: The Cohen's d family. *The Quantitative Methods for Psychology*, 14(4):242–265.

Dani Gunawan, CA Sembiring, and Mohammad Andri Budiman. 2018. The implementation of cosine similarity to calculate text relevance between two documents. In *Journal of physics: conference series*, volume 978, page 012120. IOP Publishing.

Andy Halterman. 2020. Extracting political events from text using syntax and semantics. *Technical report MIT*.

Hans Ole Hatzel and Chris Biemann. 2023. Narrative cloze as a training objective: Towards modeling stories using narrative chain embeddings. *Proceedings of the The 5th Workshop on Narrative Understanding*.

Mourad Heddaya, Qingcheng Zeng, Alexander Zentefis, Rob Voigt, and Chenhao Tan. 2024. Causal micronarratives. In *Proceedings of the 6th Workshop on Narrative Understanding*, pages 67–84.

Harshad Hegde, Jennifer Vendetti, Damien Goutte-Gattat, J Harry Caufield, John B Graybeal, Nomi L Harris, Naouel Karam, Christian Kindermann, Nicolas Matentzoglu, James A Overton, Mark A Musen, and Christopher J Mungall. 2025. A change language for ontologies and knowledge graphs. *Database: The Journal of Biological Databases and Curation*, 2025:baae133.

Maria Hellman. 2024. Narrative analysis and framing analysis of disinformation. In *Security, Disinformation and Harmful Narratives: RT and Sputnik News Coverage about Sweden*, pages 101–121. Springer.

Hansi Hettiarachchi, Mariam Adedoyin-Olowe, Jagdev Bhogal, and Mohamed Medhat Gaber. 2021. Daai at case 2021 task 1: Transformer-based multilingual socio-political and crisis event detection. *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*.

Ella Hofmann-Coyle, Mayank Kulkarni, Lingjue Xie, Mounica Maddela, and Daniel Preoţiuc-Pietro. 2022. Extractive entity-centric summarization as sentence selection using bi-encoders. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 326–333.

Ali Hürriyetoğlu, Surendrabikram Thapa, Gökçe Uludoğan, Somaiyeh Dehghan, and Hristo Tanev. 2024. A concise report of the 7th workshop on challenges and applications of automated extraction of socio-political events from text. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 248–255.

Brian Felipe Keith Norambuena, Tanushree Mitra, and Chris North. 2023. A survey on event-based news narrative extraction. *ACM Computing Surveys*, 55(14s):1–39.

Yuanyuan Lei and Ruihong Huang. 2023. Identifying conspiracy theories news based on event relation graph. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9811–9822.

João A Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. 2024. Euvsdisinfo: A dataset for multilingual detection of pro-Kremlin disinformation in news articles. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5380–5384.

Chang Li. 2021. *Improving Stance and Bias Detection in Text by Modeling Social Context*. Ph.D. thesis, Purdue University.

Haochen Li and Di Geng. 2024. Prompt-based graph model for joint liberal event extraction and event schema induction. *arXiv preprint arXiv:2403.12526*.

Jin Li, Ruifan Yuan, Yu Tian, and Jingsong Li. 2024a. Towards instruction-tuned verification for improving biomedical information extraction with large language models. *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 6685–6692.

Sha Li, Revanth Gangi Reddy, Khanh Nguyen, Qingyun Wang, Yi Fung, Chi Han, Jiawei Han, Kartik Natarajan, Clare Voss, and Heng Ji. 2024b. Schema-guided culture-aware complex event simulation with multi-agent role-play. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 372–381.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ben Liu, Jihai Zhang, Fangquan Lin, Cheng Yang, Min Peng, and Wotao Yin. 2025. Symagent: A neural-symbolic self-learning agent framework for complex reasoning over knowledge graphs. In *Proceedings of the ACM on Web Conference 2025*, pages 98–108.

Zhengdong Lu, Ziqian Zeng, Jianwei Wang, Hanlin Wang, Weikai Lu, and Huiping Zhuang. 2024. Zero-shot event argument extraction by disentangling trigger from argument and role. *International Journal of Machine Learning and Cybernetics*, pages 1–19.

Nick Mckenna, Tianyi Li, Liang Cheng, Mohammad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2758–2774.

Derek Miller. 2019. Leveraging BERT for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Cedric Möller, Jens Lehmann, and Ricardo Usbeck. 2022. Survey on english entity linking on wikidata: Datasets and approaches. *Semantic Web*, 13(6):925–966.

Gireen Naidu, Tranos Zuva, and Elias Mmbongeni Sibanda. 2023. A review of evaluation metrics in machine learning algorithms. In *Computer science on-line conference*, pages 15–25. Springer.

Brian Felipe Keith Norambuena and Tanushree Mitra. 2021. Narrative maps. *Proceedings of the ACM on Human-Computer Interaction*, 4:1 – 33.

Rrubaa Panchendrarajan and Arkaitz Zubiaga. 2024. Synergizing machine learning & symbolic methods: A survey on hybrid approaches to natural language processing. *Expert Systems with Applications*, 251:124097.

Armin Pournaki and Tom Willaert. 2024. A graph-based approach to extracting narrative signals from public discourse. *arXiv preprint arXiv:2411.00702*.

Oona Rainio, Jarmo Teuho, and Riku Klén. 2024. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1):6086.

Adrian Rauchfleisch and Andreas Jungherr. 2024. Blame and obligation: The importance of libertarianism and political orientation in the public assessment of disinformation in the United States. *Policy & Internet*, 16(4):801–817.

Denise Rey and Markus Neuhäuser. 2011. Wilcoxon-signed-rank test. In *International encyclopedia of statistical science*, pages 1658–1659. Springer.

Angelika Romanou, Syrielle Montariol, Debjit Paul, Léo Laugier, Karl Aberer, and Antoine Bosselut. 2023. Crab: Assessing the strength of causal relationships between real-world events. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15198–15216.

Brenda Santana, Ricardo Campos, Evelin Amorim, Alípio Jorge, Purificação Silvano, and Sérgio Nunes. 2023. A survey on narrative extraction from textual data. *Artificial Intelligence Review*, 56(8):8393–8435.

Cristian Santini. 2024. Combining language models for knowledge extraction from Italian TEI editions. *Frontiers in Computer Science*, 6:1472512.

Wenbo Shang and Xin Huang. 2024. A survey of large language models on generative graph analytics: Query, learning, and applications. *arXiv preprint arXiv:2404.14809*.

Zhu Shen, Ambarish Chattopadhyay, Yuzhou Lin, and Jose R Zubizarreta. 2024. An anatomy of event studies: hypothetical experiments, exact decomposition, and weighting diagnostics. *arXiv preprint arXiv:2410.17399*.

Jialong Tang, Hongyu Lin, M. Liao, Yaojie Lu, Xianpei Han, Le Sun, Weijian Xie, and Jin Xu. 2021. From discourse to narrative: Knowledge projection for event relation extraction. *ArXiv*, abs/2106.08629.

Nikos Voskarides, Edgar Meij, Sabrina Sauer, and Maarten de Rijke. 2021. News article retrieval in context for event-centric narrative creation. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 103–112.

Guan Wang, Rebecca Frederick, Jinglong Duan, William BL Wong, Verica Rupar, Weihua Li, and Quan Bai. 2025. Detecting misinformation through framing theory: the frame element-based model. *Journal of Computational Social Science*, 8(3):72.

Shaofei Wang. 2025. Enhancing in-context learning of large language models for knowledge graph reasoning via rule-and-reinforce selected triples. *Applied Sciences*, 15(3):1088.

Zihao Wang, Hang Yin, and Yangqiu Song. 2022. Logical queries on knowledge graphs: emerging interface of incomplete relational data. *IEEE Data Eng. Bull.*, 45(4):3–18.

Yilin Wen, Zifeng Wang, and Jimeng Sun. 2024. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10370–10388.

Tzu-Tsung Wong and Po-Yang Yeh. 2019. Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1586–1594.

Peiyi Yang, Bin Song, and Zhiyong Zhang. 2024. Research on knowledge graph construction methods for news domain. *Academic Journal of Science and Technology*, 11(1):58–64.

Hao Zhang, Frank Boons, and Riza Batista-Navarro. 2019. Whose story is it anyway? Automatic extraction of accounts from news articles. *Information processing & management*, 56(5):1837–1848.

Tong Zhou, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Cogmg: Collaborative augmentation between large language model and knowledge graph. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 365–373.

Shenzhe Zhu. 2024. Exploring knowledge graph-based neural-symbolic system from application perspective. *arXiv e-prints*, pages arXiv–2405.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2024. LLMs for knowledge graph construction and reasoning: recent capabilities and future opportunities. *World Wide Web*, 27(5):58.

Ling Zhuang, Hao Fei, and Po Hu. 2023. Knowledge-enhanced event relation extraction via event ontology prompt. *Information Fusion*, 100:101919.

A Examples of News Article Summaries

In this section a few examples of news article summaries are presented together with their ROUGE scores. The summary extracted with the method that achieved higher ROUGE score, which reflects higher overlap with original text, was used in the experiments (Table 12).

Title	LexRank summary	BERT extractive summary	ROUGE (LexRank)	ROUGE (BERT)
EU supports Zelensky's peace plan – Michel	Ukraine is The European Union and The European Union is Ukraine. In Africa, in Asia, China and in Latin America, in every meeting with foreign leaders, support for Ukraine is the first issue The European Union raise, European Council President Charles Michel said. The damages caused by Russia to Ukraine must be compensated and all those – without exception – responsible must be and will be held accountable, European Council President Charles Michel said. The European Union have frozen over EUR 300 billion of Russian assets. Since day one of Russia's attack, Europeans have understood that Russia's attack is more than an attack on Ukraine. The European Union is more united than ever.	The European Union supports the Peace Plan proposed by President of Ukraine Volodymyr Zelensky, European Council President Charles Michel has said. You this goal an end to this war of aggression and The European Union shares this goal with you. The damages caused by Russia to Ukraine must be compensated and all those – without exception – responsible must be and will be held accountable, European Council President Charles Michel said. European Council President Charles Michel am strongly convinced over EUR 300 billion of Russian assets should and will be used for Ukraine's reconstruction. Since day one of Russia's attack, Europeans have understood that Russia's attack is more than an attack on Ukraine.	0.291	**0.292**

Title	LexRank summary	BERT extractive summary	ROUGE (LexRank)	ROUGE (BERT)
International Space Station preparing for 'space taxis'	Two astronauts will prepare the International Space Station for two new docks for the spacecraft that will ferry astronauts and gear to and from the International Space Station. Two new docks for the spacecraft that will ferry astronauts and gear to and from the station will allow both Boeing CST's and SpaceX's Dragon spacecraft to dock on the American side of the ISS, even though the two companies' ships are designed differently. The astronauts will be setting up cables on the first two walks and antennae on the third walk scheduled for March 1. It will be the first time vessels have docked on the American side of the ISS since NASA ended the Space Shuttle program. NASA awarded Boeing a $4.2 billion contract in September to develop a transportation capable of carrying human passengers.	A spacewalk outside the ISS planned for Friday was postponed by a day, after "added analysis of spacesuits" the astronauts will wear, according to a NASA statement. NASA confirmed it needed more time to look at suits but did not give additional information. Boeing is building two new docks for the spacecraft that will ferry astronauts and gear to and from the station. SpaceX will carry two new docks to the ISS on cargo mission CRS-7. If all goes according to plan, it will be the first time vessels have docked on the American side of the ISS since NASA ended the Shuttle program. It will also allow NASA to increase crew size and scientific research.	**0.622**	0.59

Title	**LexRank summary**	**BERT extractive summary**	**ROUGE (LexRank)**	**ROUGE (BERT)**
The Maduro Diet: Food v. Freedom in Venezuela	Nicolas Maduro's regime has been unable to control inflation, which has disintegrated Venezuela's economy. As a result, many poor Venezuelans are now dependent on CLAP deliveries to put food on the table. A politically motivated food system called CLAP has become an essential part of nearly half the population's diet. Those overseeing the program earn a 200% profit per box. Though the regime aims to reduce dependence on imports, 90% of CLAP boxes come from Mexico. The international community should link external pressure with renewed political opposition to bring about a democratic transition.	As Venezuela's humanitarian crisis deepens, Maduro's regime has exploited it to tighten political control. Poverty has risen from 55% in 1998 to nearly 90% today. Though the regime claims to reduce food import dependence, 90% of CLAP boxes come from Mexico. The weaponization of the CLAP program is a sign of a desperate regime. The use of cryptocurrency-based aid models like EatBCH may help.	**0.294**	0.212

Table 12: Examples of news article summaries

B Statistics of the Final Dataset

This section presents a basic analysis of the final dataset used for experiments. Table 13 presents article distribution per year in the final dataset. For trustworthy articles, 2020 had the largest number of articles (109), while for disinformation the articles were distributed more equally, with 2019 and 2020 having the largest numbers (46 each).

Year	Disinformation	Trustworthy
2015	18	3
2016	26	9
2017	29	14
2018	43	24
2019	46	52
2020	46	109
2021	28	22
2022	43	52
2023	28	17

Table 13: Article distribution 2015-2023.

Table 14 lists publishers with the highest numbers of articles in the final dataset, where for disinformation the main publisher was RT (166 articles) and for trustworthy news the articles were more equally distributed with BBC having the largest number of articles (28).

Publishers	
Disinformation	**Trustworthy**
RT (166)	BBC (28)
TASS (51)	Polygraph.info (25)
Sputnik (15)	The Guardian (24)
Geopolitica.ru (12)	Radio Free Europe (11)
Global research (11)	DW.com (10)

Table 14: Publishers with the largest number of articles (number in brackets) in the final dataset.

Similarities and differences in topics across disinformation and trustworthy articles in the final dataset can be seen in word clouds (Fig. 1) representing the most frequent 25 words, excluding stopwords, of each part, with ’russian’ being the most frequent word word in both groups of articles. Also, in both groups Russian-Ukrainian war related word forms are prominent, with ’said’ being the most frequent word form. Furthermore, in trustworthy articles ’coronavirus’ is among the most frequent word forms, while in disinformation articles this word form is not among most frequent ones, indicating difference in topic coverage.

C Prompts Used for LLM-only and Augmented Extractions of Narrative Elements

For LLM-only extraction of narrative elements, the following prompt was used:

```
prompt_text = (
"Extract the following narrative
elements from the news article:\n"
"- Characters (Who are the key actors?)\n"
"- Events (What happened?)\n"
"- Causal Links (Why did it happen?)\n"
"- Framing:\n"
"    - Actor Framing: How are key actors
(individuals, organizations) portrayed?
(e.g., hero, villain, expert, victim)\n"
"    - Event Framing: How is the event
presented? (e.g., crisis, scandal,
breakthrough, tragedy)\n\n"
f"News summary: {best_summary}")
```

For Augmented pipeline, the prompt below was used:

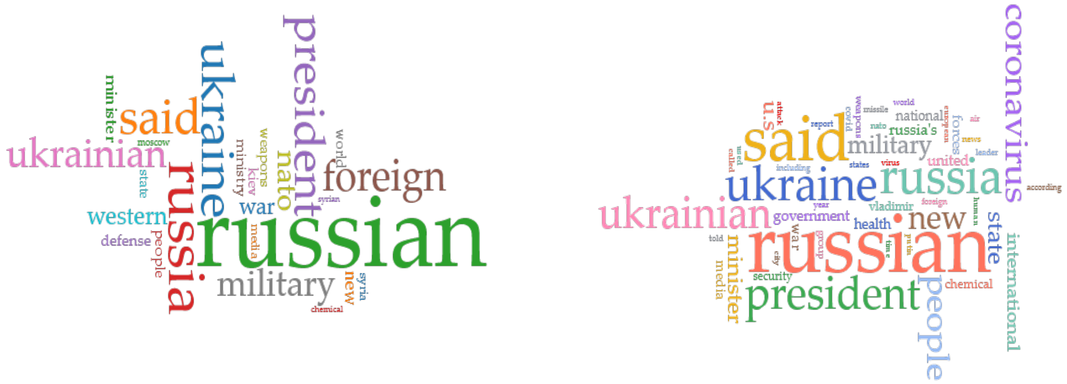
```
prompt_text = (
"Extract the following narrative elements
from the news article:\n"
"- Characters (Who are the key actors?)\n"
"- Events (What happened?)\n"
"- Causal Links (Why did it happen?)\n"
"- Framing:\n"
"    - Actor Framing: How are key actors
portrayed? (e.g., hero, villain)\n"
"    - Event Framing: How is the event
presented? (e.g., crisis, breakthrough)\n\n"
f"News summary:\n{best_summary}\n\n"
f"Knowledge Graph context:\n{kg_context}")
```

D REBEL Relation Sets/Mappings

The relations of REBEL schema that imply framing, events and causality were mapped and inferred manually:

Relations suggesting events or time-anchored nodes:

- Participation-based: "participant_in", "participant", "winner", "candidate", "candidacy_in_election".
- Achievement or outcome: "award_received", "notable_work", "nominated_for".
- Conflict or disruption: "conflict", "significant_event", "victory", "replaced_by", "replaces".
- Time anchors: "inception", "start_time", "end_time", "point_in_time", "publication_date", "date_of_birth", "date_of_death", "work_period_(start)", "work_period_start".



(a) Most frequent word forms in disinformation articles

(b) Most frequent word forms in trustworthy articles

Figure 1: Most frequent word forms

- Organizational change: "founded_by", "dissolved_abolished_or_demolished_date", "location_of_formation", "established_date".
- Media & Creative: "original_broadcaster", "composer", "director", "producer", "production_company"

Relations suggesting causality:

- Explicit causality: "has_cause", "causes", "has_effect", "influenced_by", "leads_to", "responds_with".
- Implied causality: "participant_in", "participant", "winner", "founded_by", "replaced_by", "replaces", "award_received", "influenced_by", "significant_event", "conflict", "start_time", "end_time"

Relations suggesting framing:

- Explicit framing:
  - "described_as": "is portrayed as"
  - "depicts": "is represented as"
  - "notable_work": "is known for"
  - "award_received": "is recognized for"
  - "member_of_political_party": "is affiliated with"
  - "replaced_by": "is portrayed as outdated or replaced by"
  - "founded_by": "was founded by"
- Causal interpretation framing:
  - "has_cause": "is caused by"
  - "causes": "causes",
  - "has_effect": "results in"
  - "influenced_by": "is influenced by"
  - "leads_to": "leads to"
  - "responds_with": "is responded to with"
  - "participant_in": "participated in"
  - "participant": "participated in"
  - "winner": "wins due to"
  - "candidate": "candidacy in"
  - "significant_event": "is triggered by"
  - "conflict": "is involved in conflict"
  - "start_time": "starts at"
  - "end_time": "ends at"

E Downstream Classification Setup

The downstream classification setup for extrinsic evaluation consisted of the following:

- For each pipeline, a feature matrix was formed.
- Feature sets:
  - LLM-only: counts of narrative elements and values of embedding-based coherence measure produced by LLM-only pipeline from extractive summaries of news articles.
  - KG-only: counts of narrative elements and values of graph-based coherence measures (density, avg. path length, largest component ratio) produced by KG-only pipeline from extractive summaries of news articles.

  - Augmented: counts of narrative elements and values of embedding-based coherence measure produced by Augmented pipeline from extractive summaries of news articles.

- For classification, a scikit-learn pipeline was used:

  StandardScaler() → LogisticRegression(solver='liblinear', penalty='l2', C=1.0, max_iter=100) (defaults shown).

- Evaluation was performed with StratifiedKFold (5 and 10 splits, shuffle=True, random_state=42) using F1 as the primary metric via cross_val_score. Mean ± SD across folds are reported as well. Significance is assessed on fold-wise F1 with paired t and Wilcoxon signed-rank tests. Effect sizes were evaluated with Cohen's d.

Abstractive Event Analysis of Armed Conflicts: Introducing the UCDP-AEC Dataset

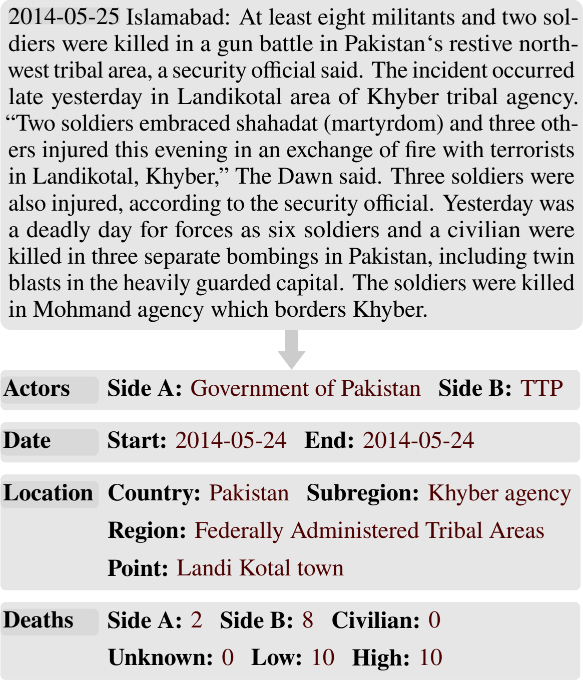
Étienne Simon[*1] and **Helene Bøsei Olsen**[*1] and **Ramón Carreño**[1] and **Rahul Mishra**[1]
Nikolay Arefyev[1] and **Mert Can Yilmaz**[2] and **Lilja Øvrelid**[1] and **Erik Velldal**[1]
[1] University of Oslo, [2] Uppsala University

Abstract

This paper introduces a new dataset of document-level event annotations in the domain of armed conflict. By augmenting the event database from the Uppsala Conflict Data Program (UCDP) with source documents identified in public web archives, we create the *UCDP Abstractive Event analysis Corpus* (UCDP-AEC). While a large part of research on information extraction is focused on extracting text spans, real-world use cases often require inferring more high-level information that is not necessarily explicitly mentioned in texts. UCDP-AEC differs from traditional event extraction datasets in that the document-level annotations do not correspond to mere text spans of the input, but capture expert-interpreted and often implicit information. With more than 10 000 documents, UCDP-AEC is of comparable size to the largest human-annotated traditional event extraction datasets. We also report preliminary experimental results for various generative approaches, by fine-tuning both decoder models and existing event argument extraction models that require minimal adaptation to our abstractive formulation of the task.

1 Introduction

For several decades, the Uppsala Conflict Data Program (UCDP) has compiled and released datasets on quantitative armed conflict event data (Sundberg and Melander, 2013; Davies et al., 2024), designed primarily to enable research on the causes, dynamics, and consequences of organised violence, as well as supporting research projects such as predicting conflict escalation (Hegre et al., 2022). UCDP's event database, the Georeferenced Event Dataset (GED), encodes rich information about incidents related to organised violence, including casualty counts, involved actors, location, time, and more. An example of (a relevant subset of) the data fields and corresponding values for a given input document is shown in Figure 1. Carried out through manual coding by human experts, the process is highly resource-demanding and time-consuming, making it challenging to scale to larger data sources and update in near real-time. Moreover, the continual exposure to accounts of violence can cause emotional and psychological distress for those involved in manual coding. For several reasons, automated machine-coding of events could offer valuable assistance in the annotation process.



2014-05-25 Islamabad: At least eight militants and two soldiers were killed in a gun battle in Pakistan‘s restive northwest tribal area, a security official said. The incident occurred late yesterday in Landikotal area of Khyber tribal agency. “Two soldiers embraced shahadat (martyrdom) and three others injured this evening in an exchange of fire with terrorists in Landikotal, Khyber,” The Dawn said. Three soldiers were also injured, according to the security official. Yesterday was a deadly day for forces as six soldiers and a civilian were killed in three separate bombings in Pakistan, including twin blasts in the heavily guarded capital. The soldiers were killed in Mohmand agency which borders Khyber.

Actors **Side A:** Government of Pakistan **Side B:** TTP

Date **Start:** 2014-05-24 **End:** 2014-05-24

Location **Country:** Pakistan **Subregion:** Khyber agency
Region: Federally Administered Tribal Areas
Point: Landi Kotal town

Deaths **Side A:** 2 **Side B:** 8 **Civilian:** 0
Unknown: 0 **Low:** 10 **High:** 10

Figure 1: Sample document with UCDP annotations. The document reports on an armed conflict event and includes the publication date and location. The annotations specify the two conflict actors, the start and end dates of the event, four levels of location detail, and death counts for each side, civilians and unknown individuals. Estimated total deaths are given as low (conservative) and high (maximum reliable) figures.

The UCDP GED example above illustrates a

---

[*] Equal contribution.

Proceedings of the 5th Workshop on Computational Linguistics for the Political and Social Sciences (CPSS), pages 98–113
September 10-11, 2025 ©2025 Association for Computational Linguistics

more general challenge acknowledged in other recent work on event extraction (EE) and information extraction (IE) for real-world applications, which is the need to move beyond span-based annotations. For example, Olsen et al. (2024) refer to an ‘abstraction gap’ when discussing the differences between the text-bound annotations common in most NLP datasets for EE versus the higher-level information documented in socio-political event databases. In such databases, coders aim to represent what actually happened in the real world, drawing on contextual understanding and domain expertise. As demonstrated in Figure 1, important event information often can not be found as a single continuous text span. Instead they must often be inferred from indirect references (e.g., “soldiers” or “militants” as proxies for specific groups), resolved through temporal reasoning (e.g., interpreting “yesterday” relative to the publication date of the source document), aggregated from scattered mentions (e.g., casualty counts reported across multiple sentences), or produced through numerical reasoning (e.g., combining the casualty counts across parties). Locations and actors can be described at varying levels of granularity and may require entity-linking to canonical forms, and the dynamic nature of conflicts means new actors and conflicts continually emerge, introducing important temporal variations. On the basis of considerations like this, Simon et al. (2024) advocate for a shift from an “extractive” to an “abstractive” view of the EE task, proposing to use generative approaches for capturing more higher-level event information. Similarly, Sharif et al. (2024) also show how event information can be scattered across a document or only implicitly stated, and propose a generative formulation of event argument extraction.

While the recognition of the challenges associated with real-world applications is growing, the majority of traditional event extraction datasets within Natural Language Processing (NLP) are annotated in the extractive paradigm, capturing only text-bound facts (Doddington et al., 2004; Song et al., 2015; Li et al., 2021). On the other hand, while there is an abundance of high-quality event databases developed within social and political sciences (Sundberg and Melander, 2013; Raleigh et al., 2010; Chenoweth et al., 2019; Salehyan et al., 2012; Turchin, 2012), the source documents are generally not made available (Olsen et al., 2024), representing a missed opportunity for training machine-learning models for domain-specific applications. As a case in point, the event annotations of the UCDP GED are publicly available, but the corresponding source documents used by the human coders are not licensed for redistribution in the original database.

Contributions To bridge NLP modelling with socio-political event coding, we introduce the new dataset UCDP-AEC, a conflict event dataset for abstractive event analysis built from UCDP GED, augmented with pointers to the underlying source texts.[1] The dataset covers 11 363 documents describing armed conflict events and links expert-annotated event records to their corresponding source documents. UCDP-AEC focuses on a single event type, and every document in the dataset contains a single, relevant event of the same type, where each event is annotated with a fixed set of 14 fields. Reflecting the complexities of real-world event reports, the annotations often demand numerical and temporal reasoning, world- and domain knowledge, contextual integration, and entity linking to canonical forms. We refer to this challenging task of structured prediction as abstractive event analysis.

We show how a substantial subset of the documents from the UCDP GED database can be identified and retrieved from the public web archives of HPLT v2 (High Performance Language Technologies; Burchell et al., 2025), and that distributing the corresponding document IDs provides a way to establish an open dataset for event analysis that can be used by the NLP community, ultimately also benefitting the peace and conflict studies. Finally, we report on the first experimental results for a suite of generative approaches on the new dataset.

2 Related Work

Relation to other NLP tasks Automatically extracting UCDP-AEC event structures combines elements from multiple NLP tasks involving representing and understanding event-centric information in text. The task includes aspects of temporal reasoning, text summarisation, and entity-linking, but is best understood as a form of event extraction (or perhaps more narrowly the associated subtask of event argument extraction) though in a way that diverges from typical NLP formulations.

While event extraction is a widely studied task in NLP, the current annotation paradigm and bench-

---

[1] The dataset alongside evaluation scripts and baseline models are available at `https://github.com/ltgoslo/ucdp-aec`.

mark datasets do not correspond well to the types of events typically encoded in socio-political event databases. Even though both tasks aim to derive structured event information from unstructured text, existing NLP datasets typically contain sentence-level event information that corresponds to substrings in the input text (Doddington et al., 2004; Song et al., 2015), while socio-political event recordings typically capture higher-level event information inferred from the entire document, possibly in combination with domain knowledge. For instance, in the news article shown in Figure 1, Side B is described in the text as "militants" and "terrorists". However, a socio-political event database is expected to convey the specific group involved ("TTP" in this case) based on contextual information such as time and location.

The datasets from the Automatic Content Extraction (ACE; Doddington et al., 2004) program have been highly influential in the development and evaluation of event extraction systems in NLP, inspiring further development of several datasets with more detailed processing of entities and events in the ERE (Entities, Relations and Events) dataset (Song et al., 2015), document-level event extraction from Wikipedia articles in the WikiEvents dataset (Li et al., 2021), from news articles referenced in Wikipedia articles in the DocEE dataset (Tong et al., 2022), and domain-specific content (Sun et al., 2022; Satyapanich et al., 2020).

Some recent works have taken a step towards removing the reliance on surface forms for event extraction. For the task of Chinese financial event extraction, Zheng et al. (2019) do not extract trigger words, but still rely on text spans for argument extraction. More recently, Sharif et al. (2024) challenge the traditional span-based approach and introduce DiscourseEE, an online health discourse dataset annotated with explicit, implicit, and scattered arguments. Apart from DiscourseEE, a related work to ours is the early version of the task introduced at the Message Understanding Conferences (MUC; Sundheim, 1992),with the most well-known being the MUC-4 dataset, which is annotated with rich, high-level, and fine-grained information in the form of event templates at the document level. Recent work has renewed focus on the MUC-4 data, however, existing work is largely limited to argument roles that correspond directly to explicit text spans from the source document (Du et al., 2021a,b; Gantt et al., 2024).

Concurrent to our work Semnani et al. (2025) introduce LEMONADE, a large-scale multilingual dataset for abstractive event extraction based on a re-annotated subset of the ACLED database (Raleigh et al., 2010). It spans 25 closely related socio-political event types over a 13-month period and includes an abstractive entity linking task to a curated domain-specific entity database. Each event type is associated with a specific schema, which supports a variety of argument types, including categorical fields, booleans, and integers, with many roles limited to values in a pre-defined database, and some re-annotated to only reflect information explicitly present in the text. In contrast, UCDP-AEC focuses on a single domain, fatal armed conflicts, grounded in the expert coding of UCDP GED, with a fixed 14-field schema that is always filled, often requiring inference over implicit or scattered evidence. UCDP-AEC contains multiple open domain roles, such as exact event dates or fine-grained location details, without requiring that the gold value appear verbatim in the document.

Largely independently of the NLP community, the political science community has seen the development of their own event extraction systems (Schrodt et al., 1994; Norris et al., 2017; Halterman et al., 2023). However, these systems tend to rely on older rule-based models prone to overcounting events and producing numerous false positives, which has limited their use in political science research (Raleigh et al., 2023). Not relying on learned models and annotated datasets also makes most systems less flexible and adaptable. Some initial work making training data available is worth mentioning though, such as GLOCON (Global Contentious Politics Dataset; Duruşan et al., 2022) and CEHA (Conflict Events in the Horn of Africa region; Bai et al., 2025), with the latter reaching 500 events, but with annotations for event detection and classification only.

Generative approaches to event extraction In recent years, the NLP field has seen a transition from traditional sequence labelling approaches for event extraction (Wadden et al., 2019; Nguyen et al., 2016) towards generative methods where the task is framed as a structured text generation task (Lu et al., 2021; Li et al., 2021). Generative methods present a promising path to move beyond span-based annotations by allowing models to generate structured event representations directly from text, and may be suitable for tasks requiring inference from under-specified information and long document contexts, such as for UCDP-AEC.

However, the current landscape of generative approaches to event analysis still remains closely tied to the extractive paradigm (Simon et al., 2024). As the architecture and training objective of generative models are intended for producing free-form text, they provide a less natural fit to the extractive setting, where the goal is to identify and reproduce exact text spans from the source text. As a consequence, implementations of generative event extraction models often include additional constraints to ensure that the generated string appears in the input, such as constraint decoding in the Text2Event model (Lu et al., 2021), or the pointer mechanism in the BART-Gen model (Li et al., 2021).

Recently, several decoder-only models have been developed for generative event extraction, such as DeepStruct (Wang et al., 2022), InstructUIE (Wang et al., 2023), YAYI-UIE (Xiao et al., 2023), a Baichuan2 model (Yang et al., 2023), and LLMEE (Chen et al., 2024). Notably, except for DeepStruct and YAYI-UIE, most works, including LLMEE, do not involve any task-specific fine-tuning, relying solely on in-context learning. While these models, in addition to some encoder-decoder models such as DEGREE (Hsu et al., 2022), do not explicitly constrain their outputs to input tokens during prediction, they remain in the extractive paradigm through evaluation, where the predictions are converted into text spans for evaluation against benchmarks such as ACE (Doddington et al., 2004).

3 The UCDP GED Conflict Database

An event in the UCDP GED database is defined as an incident where armed force was used by an organised actor resulting in at least one direct death at a specific location and date. In this section, we provide some background on the UCDP GED annotations, starting with the annotation process itself, followed by a description of the relevant data fields that we include in the derived UCDP-AEC dataset described in Section 4.

3.1 Annotation Process

The UCDP GED database follows a rigorous, multi-step process to document each instance of fatal organised violence as a distinct event (Sundberg and Melander, 2013). Data collection begins with global search queries in the Dow Jones Factiva aggregator, yielding tens of thousands of news reports annually. These reports are then supplemented with local and social media sources, as well as information from NGOs and international organisations. Reports are meticulously examined by a team of around 15 analysts, who hold advanced degrees in peace and conflict studies or related fields and possess extensive regional and conflict-specific expertise. They identify and code events according to strict methodological criteria, assigning detailed spatial, temporal, and actor identifiers (Högbladh, 2023). The individual event entries are subsequently categorised into three mutually exclusive types of violence: state-based, non-state, or one-sided.

A candidate event is only included in the final event database after several stages of validation, including both manual review and automated consistency tests. In sum, this comprehensive coding approach aims to provide a high level of accuracy and reliability in capturing global patterns of organised violence, and updated versions of UCDP GED are published annually.[2] The current full GED contains more than 500 000 events with a good coverage of all world conflicts from 1989 onwards.

3.2 Data Fields

The UCDP GED database contains a wide range of different data fields, which, beyond event information, include source metadata, estimates of annotation reliability, identifiers linking to other databases, and more. In this work, however, we limit the description to the 14 fields selected for the new UCDP-AEC dataset as shown in Figure 1 and in Appendix C.[3] We group the event fields into the following four categories, each with associated challenges.

Actors The two fields “Side A” and “Side B” corresponds to the two parties of a conflict. “Side A” is always an organised actor, while “Side B” can be either an organised actor or “Civilians” when the first side kills people indiscriminately. All mentions of the same actor are labelled with a unique canonical name, rather than how they are referred to in the text.

Dates The two fields “Start Date” and “End Date” reference a time range during which the event occurred. As UCDP aims to record events with a day-level precision, the start and end dates will typ-

---

[2] `https://ucdp.uu.se/downloads/`

[3] For the complete list and description of the fields in UCDP GED, see Högbladh (2023).

ically be identical for most events.[4] Both relative time references – also known as temporal deixis – and the publication date and time are used to identify the start and end date of an event, as shown in Figure 1.

Locations The location of the event is described by four fields with an increasing level of precision. The first field, country, is always provided and indicates where the event took place. Each country is divided into region and subregion, which may be left empty if the location does not align with administrative areas. The point field refers to a city-level location, such as a town or a specific site, such as "Jabal al Akrad mountain", and may also be empty if the location is unclear.

Deaths Six distinct fields are used to describe the casualties associated with the conflict event. Specifically, the fields represent the number of casualties on each side of the conflict, the number of civilian casualties, and the number of deaths that cannot be attributed to any of those categories. In the case of one-sided violence – when "Side B" is "Civilians" – the number of civilian casualties is reported in the field "Deaths Civilians". Uncertainty in casualty reporting is reflected in the fields "Deaths Low" and "Deaths High", representing the most conservative estimate and the highest reliable estimate, respectively, for the total number of fatalities reported in the source text.

4 The UCDP-AEC Dataset

In this section, we describe the construction of the new UCDP-AEC dataset designed for machine learning applications based on the socio-political database UCDP GED. We begin by outlining the event filtering process, which ensures a one-to-one correspondence between events and documents. We then outline how we used publicly available web archives to make the data available. Finally, we present some descriptive statistics of UCDP-AEC, followed by a discussion of relevant challenges.

4.1 Event Selection

Since the UCDP GED is annotated for political science research rather than machine learning applications, some filtering was necessary to adapt the data for automated event analysis. In particular, some events are referenced by multiple source articles, some of which reference multiple events. In practice, the annotators are aware of which parts of a document have already been coded in the database, ensuring that there is no confusion. However, from an NLP perspective, this would require modelling the problem as mapping a set of documents to a set of events without a simple one-to-one correspondence between them. To simplify the problem, we only include events described by a single source and sources that describe a single event. Furthermore, because of the limited amount of non-English documents in the selected subset, we restricted UCDP-AEC to English-language documents. Additional details on data filtering are provided in Appendix A.

4.2 Web Archive Identification of Sources

HPLT matching The documents used to create the GED are not public. Since UCDP relies in part on large news networks for its source documents, we cannot indiscriminately share them. Instead, we distribute only the IDs of documents that can be found freely on the web. To this end, we make use of the recently released HPLT v2 dataset,[5] which consists of texts extracted from web crawls provided by two major web crawling initiatives, Common Crawl and the Internet Archive. It contains 21 billion documents released under a CC0 licence. Using the MinHash algorithm (Broder, 1997), we selected the UCDP documents having an approximate Jaccard similarity of at least 0.5 with an HPLT document.

Human evaluation of HPLT matching The search for documents in HPLT web crawls introduces some biases. In particular, the older a document is, the more likely it is to have been crawled. We recovered only 11 documents from 2023 and none from 2024, despite thousands being annotated in UCDP during those years.

Furthermore, the MinHash document matching used to replace the original UCDP documents with their HPLT version is not exact. It is not uncommon for news providers to update their articles as new information becomes available. These different versions of the same article have a high syntactic similarity – as measured by their Jaccard index – but can convey different event information – for example, when the number of deaths is updated. To assess the quality of the HPLT substitution, we performed a manual comparison of HPLT documents with their original UCDP documents.

---

[4] In some cases, the precise date remains uncertain due to vague temporal references in the text, such as "last week". In such cases, the start and end dates will differ by a week.

[5] `https://hplt-project.org/datasets/v2.0`

Split	Train	Validation	Test
Documents	10 064	651	648
Period	< 2021	2021	≥ 2022

Table 1: Number of documents and years in each split.

Three annotators assessed a total of 130 documents and found 96.9% documents were a perfect match, with 98.5% containing the same information for 3 out of the 4 categories of fields in the event records (Actors, Date, Location, Deaths).

4.3 Dataset Statistics and Analysis

UCDP-AEC consists of 11 363 document–event pairs, where each document is annotated with a single event record with the 14 fields described in Section 3. These fields vary in how constrained their value spaces are, a property we call domain openness. In this subsection, we describe the dataset splits and document length, as well as challenges associated with each field with a particular focus on how open or closed the value space is for each role.

Domain openness We define a closed-domain field as one whose admissible values make up a relatively small, and mostly fixed set that recurs frequently in the corpus (e.g., country names). In contrast, an open-domain field is characterised by a large or evolving set of admissible values, often including many items not seen during training (e.g., event-specific dates). To quantify this property, we report for each field in Table 2 the three following metrics: *value density* $\frac{|\mathcal{D}|}{|\mathcal{Y}|}$ which denotes the mean number of instances per unique value. Here, lower values indicate a more open domain. *Unique value overlap* $\frac{|\mathcal{Y}_{\text{test}} \cap \mathcal{Y}_{\text{train}}|}{|\mathcal{Y}_{\text{test}}|}$ denotes the proportion of unique test values that also occur in the training set. Finally, *instance-level overlap* $\frac{|\mathcal{D}_{\text{test}} \cap \mathcal{D}_{\text{train}}|}{|\mathcal{D}_{\text{test}}|}$ measures the proportion of test instances whose value is observed during training. We describe these metrics in more detail in Appendix B.

Standard splits To mimic the complexity of modelling a dynamic domain, where new conflicts and violent groups arise and relationships between entities change, we provide standard splits based on temporal information, considering both the occurrence and reporting times of events. Events from 2021 are taken for validation, older events are used for training, while the most recent events, from 2022 and later, are reserved for the test split. See Table 1 for the size of the respective splits.

Field	Density	Unique Overlap	Instance Overlap
Source article	1.00	0.00%	0.00%
Source date	1.76	0.00%	0.00%
Side A	27.31	71.74%	94.91%
Side B	26.24	60.24%	80.25%
Date Start	3.17	0.00%	0.00%
Date End	3.15	0.00%	0.00%
Location Country	129.12	93.18%	99.54%
Location Region	18.78	81.09%	83.33%
Location Subregion	6.49	51.30%	60.80%
Location Point	2.75	27.76%	40.74%
Deaths Side A	183.27	83.33%	99.54%
Deaths Side B	155.66	100.00%	100.00%
Deaths Civilian	195.91	94.74%	99.85%
Deaths Unknown	241.77	100.00%	100.00%
Deaths Low	113.63	90.00%	99.54%
Deaths High	97.96	90.24%	99.38%

Table 2: Per field statistics showing the value density, the unique value overlap between test and train set, and the instance-level overlap between test and train.

Document length The average length of the 11 363 documents in UCDP-AEC is 315 words (white-space separated / non-tokenised). However, there is a significant variation in document length, ranging from as few as 26 words to 1 781 words. This wide range reflects the variation in source materials used to collect armed conflict events, from short summaries to longer regional reports.

Location distribution The dataset covers events that occur in 88 countries. Figure 2 shows the temporal shifts in the geographical distribution between the splits. For example, the training set is dominated by events occurring in Syria, the validation set (2021) shows increased violence in countries such as the Philippines and Myanmar, and the test set reflects the heightened violence in Nigeria alongside the further invasion of Ukraine. With respect to openness, illustrated in Table 2, the set of possible country values is, to a large degree, limited and fixed. With only a few countries in the test set that are not present in the training set, the country field should be considered a relatively closed domain. There is, however, a relationship between location granularity and domain openness, where Subregion can be regarded as semi-open, while Region and Point tend to be highly event-specific, with a test

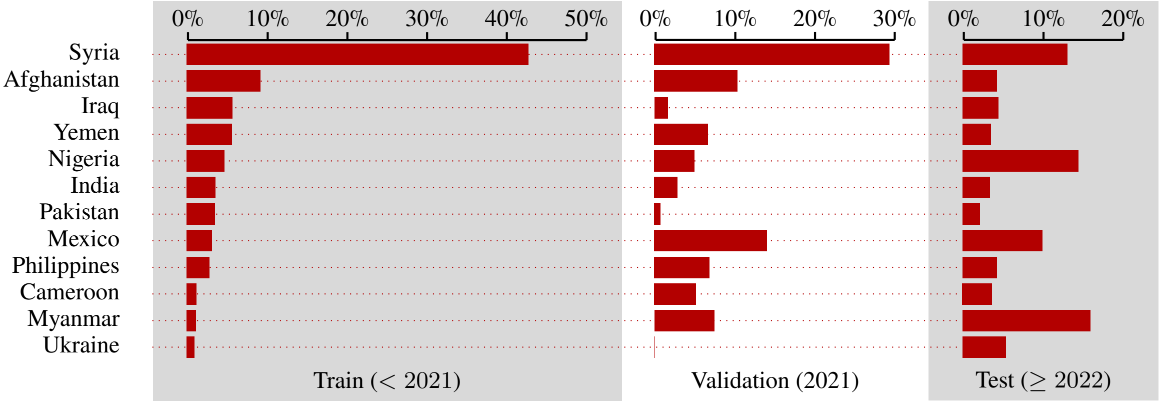


Figure 2: Repartition of events per country in each split for the 12 most eventful countries. A clear increase in conflict events can be seen in Nigeria and Myanmar, showcasing how the dataset reflects the changing dynamics of real-world conflict over time.

set containing many unseen values.

Actors The Actor roles often feature governments, organisations, or group names that are moderately recurrent, but also include many entities that appear only in specific events. Due to the dataset's temporal splits, a notable proportion of actor names in the test set do not appear in the training set, mirroring the involvement of new actors in ongoing and emerging armed conflicts. For example, the test set includes events involving the Indigenous People of Biafra (IPOB), a Nigerian separatist group that started engaging in violent attacks in 2021.[6] Several test documents labelled with IPOB contain references to members of the Eastern Security Network (ESN), which is the paramilitary wing of IPOB. Identifying the two surface forms as references to the same entities is a particularly difficult challenge.

Date Start date and end date are by design fully open (0% unique-value overlap) due to the temporal partitioning of the splits. Since UCDP-AEC only includes single-day events, 91.4% of events have matching start and end dates as expected. When the start and end dates differ, it indicates uncertainty or imprecision in the source document about the exact date of the event.

Deaths Death-related fields are largely a closed domain with high density and overlap. However, identifying the correct values from text may still be challenging. Death counts can be expressed in various ways, such as explicit numbers, ranges (e.g., "between 15 and 20"), or approximations (e.g., "dozens killed"), and require mapping to the affected individuals or parties. In the deaths category, the low and high estimates are identical in 87.1% of the cases, again indicating that the exact number is often reported. However, for 7.3% of the events, the best estimate is equal to 0 – that is the sum of deaths on side A, side B, civilians and unknown – implying that the annotator questioned the truthfulness of the reported event.

5 Experiments

This section reports on the first preliminary experiments on training and evaluating models for abstractive event analysis on UCDP-AEC.

5.1 Evaluation Metrics

Unlike classical event extraction, which only assigns roles to arguments that are explicitly mentioned in the input text, all fields in UCDP-AEC are assigned a value in all instances. Even when not explicitly stated in the document, argument values are deducted from the context. Since this is an information extraction task where the goal is to abstract away surface form differences, we evaluate model predictions using exact match accuracies for each field. This contrasts with the usual F_1 metrics reported in the classical task, where roles are not always attached to an event, as extractive models may predict excess or insufficient roles.[7] We average the accuracies of different fields over the same category. For example, we report location accuracy by averaging the four accuracies over country, region, subregion and point. Furthermore, we report

---

[6] `https://ucdp.uu.se/actor/6287`

[7] For a broader discussion on why traditional F_1 is not applicable and supplementary alternatives we considered for different categories see Appendix F.

a global accuracy aggregate across categories.[8] In the case of actors, the aggregate accuracy is invariant to a permutation of "Side A" and "Side B". If those were predicted in the opposite order of the gold annotation, the values of "Deaths Side A" and "Deaths Side B" are also inverted.

While predicting a unique symbolic representation is central to most information extraction tasks, generative models can predict semantically correct output that does not conform to the expected output (e.g. "USA" instead of "Government of United States of America"), and the abstractive nature of UCDP-AEC particularly exacerbates this. To evaluate how much of the error can be attributed to the selection of the gold surface form instead of a failure of the models' reasoning abilities, we introduce a semantic evaluation in line with Sharif et al. (2024). Given a field, all values appearing in any of the splits are embedded using BERT. When evaluating a model, its prediction for a given field is in turn embedded using BERT. The prediction is considered correct if it has the highest cosine similarity with the gold value amongst all possible values for this field. We report this value as "precision at top 1" (P@1), following the information retrieval convention. Since the identity maximises cosine similarity, the P@1 will always be greater than or equal to the accuracy, providing an upper bound to real model performance.

5.2 Baseline Models

To establish preliminary results for UCDP-AEC, we here report on experimental results for five different generative models, including both encoder-decoder and decoder-only architectures, all fine-tuned on the UCDP-AEC training set. The baselines are based on relatively small, openly available generative models to make sure the experiments are both reproducible and accessible to a broad range of researchers. These baselines are intended to explore and illustrate the specific challenges of UCDP-AEC, as well as provide a lower bound against which future work, potentially using larger and more specialised models, can be fairly compared.

In prior work, generative approaches to EE have predominantly been applied in an extractive setting, where the text generation is often constrained to the vocabulary present in the input text, as described in Section 2. Two of our baselines are based on adapting approaches originally proposed for such an extractive formulation, namely Text2Event and DEGREE, both widely used. As further detailed below, for our experiments, we make minimal modifications to the code provided with the original model papers to make them work with our abstractive formulation of the task, in which gold argument values may not appear as spans in the source text.

Text2Event (Lu et al., 2021) is using a T5 encoder-decoder architecture (Raffel et al., 2020) to translate text into S-expressions representing events. In the original setup, the model uses constrained decoding, which masks the output softmax to enforce valid S-expression structure and role names, and restricts argument generation to text spans in the input sentence. We adapt Text2Event to the abstractive setting by removing the constraint on argument generation. Additionally, we increase the input context window to 512 tokens to fit UCDP-AEC source documents, as Text2Event is originally designed for sentence-level event extraction.

DEGREE (Hsu et al., 2022) relies on the BART encoder-decoder transformer (Lewis et al., 2020) as a backbone, and the event extraction task is approached as a sequence-to-sequence problem between the input text and a natural language template to be filled. Since the UCDP-AEC events cover more fields than what DEGREE was originally designed for, we introduce a fuzzy-template matching algorithm to ensure that the model respects the template. The algorithm finds the argument assignment that minimises the Levenshtein distance between the predicted text and template, thus ignoring small variations in the natural language template output of DEGREE.

In addition, we fine-tune the Flan-T5 large model (Chung et al., 2022) using a slightly simpler formal template than that of Text2Event. Finally, we fine-tune two text-only Llama-3.2-Instruct-based models (1B and 3B, Grattafiori et al., 2024), using the Alpaca prompt format. The instruction prompt is illustrated in Appendix H.

5.3 Results

We report the performance of the baselines for each category and the aggregate accuracy in Table 3. The full per-role accuracies are given in Appendix D. We can see that for all models except Llama 3B, the date and location categories prove more challenging than the actor and deaths categories. We posit this is due to both actor and deaths having a smaller set of admissible values compared to date and location, which are closer to open-domain problems. While

---

[8] That is the average of 4 values, one for each category, not the average of the 14 fields' accuracies.

Model	**#param**	**Exact match (Accuracy %)**					**Semantic (Precision@1 %)**	
		Actors	**Date**	**Location**	**Deaths**	**Aggregate**	**Actors**	**Location**
Text2Event	223M	70.1	38.2	14.1	53.4	43.9	70.4	24.8
DEGREE	406M	73.5	64.4	**64.5**	**81.3**	70.9	**74.0**	**66.6**
Flan-T5 large	783M	**73.9**	63.0	62.8	81.0	70.2	**74.0**	65.6
Llama-3.2-Instruct	1B	71.0	70.5	59.9	76.4	69.4	71.3	61.3
Llama-3.2-Instruct	3B	73.0	**73.8**	63.9	78.5	**72.3**	73.2	65.5

Table 3: Baseline accuracies and P@1 on the UCDP-AEC test set, also showing the the number of model parameters.

the Llama 3B model demonstrates the strongest overall performance, obtaining the highest accuracy in the date category, the DEGREE and Flan-T5 large models are strong contenders, achieving the highest accuracies in the other categories.

Looking at the semantic evaluation, the P@1 follow the accuracies closely except for very low values. This seems to indicate that using single symbolic representations for the fields is not an important source of error for the most performant models.

5.4 Error-analysis

Looking into details, models struggle to predict new actors that were not observed during training. Around 5% of side A actors in the test set do not appear as actors in the training set. Out of those, Llama-3B was able to identify 25.8% correctly, while Flan-T5 large identified 32.3% correctly. A similar pattern can be observed for side B, suggesting that Flan-T5 large is better suited for previously unseen actors.

All models handle temporal uncertainty poorly, when the start and end date differ (in 8.6% of samples, usually as a result of expressions such as “over the weekend” or “last week”), date accuracy falls to 37.5% for Llama-3B and 33.8% for DEGREE, with models often failing to recognise the uncertainty and predict the same start and end date. Similarly, models tend to rely too heavily on the publication date, with a loss of performance when the event date differs from the publication date. DEGREE achieves a 35.9% accuracy in those cases, and Llama-3B a 41.5% accuracy.

6 Summary

This paper has introduced the UCDP Abstractive Event analysis Corpus – a large-scale dataset of high-quality document-level annotations of rich event representations in the domain of armed conflict, built from the UCDP GED event database and the HPLT v2 web archive. By coupling the event records to source documents identified in a public web archive, we are able to create a complete and open dataset for document-level event analysis. It makes the UCDP GED conflict annotations machine-learnable and available to the larger research community for experimentation. We believe the approach described here should also be applicable to other event databases where the underlying source documents are intact but currently not made openly available.

The annotations in UCDP-AEC have some important characteristics that set them apart from traditional event extraction datasets in NLP: Rather than sentence-level annotations tied to particular text spans with explicit information to be extracted, the document-level event representations in UCDP-AEC require inference from implied or otherwise under-specified information, possibly piecing together evidence that is scattered throughout a document. This requires a more “abstractive” approach to analysing events, which we argue may be particularly well-suited for generative approaches.

We have reported on a first suite of experiments with fine-tuning of both encoder-decoder and decoder-only models for generating event representations. While showing encouraging results and demonstrating the viability of the approach, we believe there is ample room for improvement in future work, in particular with respect to a more principled handling of event dates, and also exploring a larger space of decoder model tuning. We hope the UCDP-AEC dataset will be an important building block for future research on abstractive event analysis within both NLP and peace and conflict studies.

Limitations

While a defining feature of the UCDP-AEC dataset is that it reflects the temporal dynamics of real-world events, it offers a somewhat skewed view of global conflict dynamics compared to the underlying UCDP GED data. This misalignment is mostly an artefact of the process for deriving the dataset, as described in Section 4.2. Firstly, UCDP-AEC is a small subset of UCDP GED, representing 3.3% of the events in the database due to filtering and availability of source documents. Furthermore, because of the source quality filter – described in Appendix A – 84% of events in the dataset occurred in the last 10 years, thereby under-representing or missing many earlier conflicts.

The misalignment results in a substantial shift in the country coverage when compared to UCDP GED. One such example is Brazil, which has more than 12 000 recorded events in UCDP GED, but is represented by only four events in UCDP-AEC, 0.03% of the original. Similarly, while 40% of the UCDP-AEC events occur in Syria, only 19% of the UCDP GED events are recorded there.

The dataset we release is built from two sources: UCDP GED and HPLT. Despite providing all the code used in the creation of the dataset, since UCDP GED is partly private – the source documents cannot be shared – it is impossible to recreate the dataset without privileged access that we negotiated with UCDP. Similarly, the evaluation of the HPLT substitution described in Section 4.2 cannot be reproduced without access to the original UCDP documents.

The UCDP GED database incorporates sources from multiple languages, as sources closer to conflict zones can often provide more precise information and contribute to a broader and more representative basis for event recordings. By restricting the UCDP-AEC dataset to English-language sources, we may not only reinforce existing biases in the data, but also contribute to the ongoing over-representation of the English language in research and resources in NLP. After the event filtering and substitution with HPLT documents, described in Section 4, only a small portion of non-English sources was identified, which we decided to exclude.

All of the encoder-decoder transformers used in the baseline models have a maximum input context window of 512 tokens. Larger UCDP-AEC documents cannot fit inside this limit. Furthermore, we are using the Text2Event and DEGREE models outside of their intended extractive setting, which creates an unfair comparison with the Flan-T5 and the decoder-only baselines.

Ethics

Event data on armed conflict, especially when human lives have been lost, warrants a heightened focus on ethical implications, and particularly on the potential bias in the domain and on the bias introduced by our work. As the field advances towards event representations based on more high-level and implied information, combined with the use of large language models, there might be an increased risk of harm, with the potential of generating non-factual events.

A large proportion of the sources in the dataset consist of news articles, which present various types of bias, particularly when reporting on conflict events. Accurately documenting casualty numbers in conflict zones poses challenges due to a variety of factors, many of which are described in Seib (2021), and should be interpreted with caution. Moreover, not all armed conflicts receive media coverage, and the framing of the reported conflicts can be skewed, particularly when relying on sources in a single language (Chojnacki et al., 2012). This bias can manifest in several forms, for example, as unbalanced representations of certain actors as aggressors, the use of loaded terminology, or reliance on reports from countries with restricted freedom of the press and/or freedom of speech.

Some of the source documents in the dataset may contain names of individuals. Although the documents in the dataset consist of publicly available sources, when combined with event extraction systems, it is possible to imagine scenarios such as the surveillance of individuals, targeted monitoring of groups or other forms of malicious and harmful applications. The released dataset, however, does not include metadata or annotations related to individuals, and we suspect a greater challenge lies in the potential risks of unintentional misuse.

The substitution with HPLT document IDs was done with the intention of removing news articles covered by proprietary licences. HPLT relies on a crawler that respects the `robots.txt` convention and has a procedure to request the take down of documents. Even though HPLT has yet to receive any claims, we cannot guarantee that no such documents are included in the HPLT corpus. The HPLT project has the stated goal of providing training data for NLP applications, as such, we consider our use of their data as fair.

UCDP-AEC is released for the sole purpose of training machine learning models rather than analysing armed conflict events directly in the scope of socio-political research. The dataset presented is a simplified subset of the UCDP database and does not preserve the original distribution of events found in UCDP. For readers interested in research on socio-political events related to armed conflicts, Olsen et al. (2024) provide a comprehensive overview of available datasets.

Given the challenges with bias outlined, event extraction systems trained on UCDP-AEC are not intended to replace human experts but serve as a tool in the coding / annotation process. We stress that human evaluation is essential before automatically extracted event records can be used for further analysis or application.

The research presented in this paper, as well as the released dataset, is created to advance research aimed at better understanding why, how, and when armed conflict arises, with the main overarching goal of finding solutions to achieve a more peaceful world. We disapprove of using this dataset for research or applications that do not align with this vision and strongly discourage such usage.

References

Rui Bai, Di Lu, Shihao Ran, Elizabeth M. Olson, Hemank Lamba, Aoife Cahill, Joel Tetreault, and Alejandro Jaimes. 2025. CEHA: A dataset of conflict events in the horn of Africa. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1475–1495, Abu Dhabi, UAE. Association for Computational Linguistics.

Andrei Z Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, and 16 others. 2025. An expanded massive multilingual dataset for high-performance language technologies (hplt). *Preprint*, arXiv:2503.10267.

Ruirui Chen, Chengwei Qin, Weifeng Jiang, and Dongkyu Choi. 2024. Is a large language model a good annotator for event extraction? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17772–17780.

Erica Chenoweth, Jonathan Pinckney, and Orion A. Lewis. 2019. NAVCO 3.0 Dataset.

Sven Chojnacki, Christian Ickler, Michael Spies, and John Wiesel. 2012. Event data on armed conflict and security: New perspectives, old challenges, and some solutions. *International Interactions*, 38(4):382–401.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Shawn Davies, Garoun Engström, Therése Pettersson, and Magnus Öberg. 2024. Organized violence 1989–2023, and the prevalence of organized crime groups. *Journal of Peace Research*, 61(4):673–693.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Xinya Du, Alexander Rush, and Claire Cardie. 2021a. GRIT: Generative role-filler transformers for document-level event entity extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 634–644, Online. Association for Computational Linguistics.

Xinya Du, Alexander Rush, and Claire Cardie. 2021b. Template filling with generative transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 909–914, Online. Association for Computational Linguistics.

Fırat Duruşan, Ali Hürriyetoğlu, Erdem Yörük, Osman Mutlu, Çağrı Yoltar, Burak Gürel, and Alvaro Comin. 2022. Global contentious politics database (GLOCON) annotation manuals. *Preprint*, arXiv:2206.10299.

William Gantt, Shabnam Behzad, Hannah An, Yunmo Chen, Aaron White, Benjamin Van Durme, and Mahsa Yarmohammadi. 2024. MultiMUC: Multilingual template filling on MUC-4. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 349–368, St. Julian's, Malta. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Andrew Halterman, Philip A. Schrodt, Andreas Beger, Benjamin E. Bagozzi, and Grace I. Scarborough. 2023. Creating custom event data without dictionaries: A bag-of-tricks. *arXiv preprint arXiv:2304.01331*.

Håvard Hegre, Paola Vesco, and Michael Colaresi. 2022. Lessons from an escalation prediction competition. *International Interactions*, 48(4):521–554.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. DEGREE: A data-efficient generation-based event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1890–1908, Seattle, United States. Association for Computational Linguistics.

Stina Högbladh. 2023. UCDP georeferenced event dataset codebook version 23.1. *Department of Peace and Conflict Research*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2795–2806, Online. Association for Computational Linguistics.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309, San Diego, California. Association for Computational Linguistics.

Clayton Norris, Philip A Schrodt, and John Beieler. 2017. PETRARCH2: Another event coding program. *J. Open Source Softw.*, 2(9):133.

Helene Olsen, Étienne Simon, Erik Velldal, and Lilja Øvrelid. 2024. Socio-political events of conflict and unrest: A survey of available datasets. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 40–53, St. Julians, Malta. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Clionadh Raleigh, Roudabeh Kishi, and Andrew Linke. 2023. Political instability patterns are obscured by conflict dataset scope conditions, sources, and coding choices. *Humanities and Social Sciences Communications*, 10(1):74.

Clionadh Raleigh, Andrew Linke, Havard Hegre, and Joakim Karlsen. 2010. Introducing acled: An armed conflict location and event dataset. *Journal of Peace Research*, 47(5):651–660.

Idean Salehyan, Cullen S Hendrix, Jesse Hamner, Christina Case, Christopher Linebarger, Emily Stull, and Jennifer Williams. 2012. Social conflict in africa: A new database. *International Interactions*, 38(4):503–511.

Taneeya Satyapanich, Francis Ferraro, and Tim Finin. 2020. CASIE: Extracting cybersecurity event information from text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8749–8757.

Philip A. Schrodt, Shannon G. Davis, and Judith L. Weddle. 1994. Political science: KEDS—a program for the machine coding of event data. *Social Science Computer Review*, 12(4):561–587.

Philip Seib. 2021. *Information at war: journalism, disinformation, and modern warfare*. John Wiley & Sons.

Sina Semnani, Pingyue Zhang, Wanyue Zhai, Haozhuo Li, Ryan Beauchamp, Trey Billing, Katayoun Kishi, Manling Li, and Monica Lam. 2025. LEMONADE: A large multilingual expert-annotated abstractive event dataset for the real world. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25813–25852, Vienna, Austria. Association for Computational Linguistics.

Omar Sharif, Joseph Gatto, Madhusudan Basak, and Sarah Masud Preum. 2024. Explicit, implicit, and scattered: Revisiting event extraction to capture complex arguments. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12061–12081, Miami, Florida, USA. Association for Computational Linguistics.

Étienne Simon, Helene Olsen, Huiling You, Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2024. Generative approaches to event extraction: Survey and outlook. In *Proceedings of the Workshop on the Future of Event Detection (FuturED)*, pages 73–86, Miami, Florida, USA. Association for Computational Linguistics.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, Denver, Colorado. Association for Computational Linguistics.

Zhaoyue Sun, Jiazheng Li, Gabriele Pergola, Byron Wallace, Bino John, Nigel Greene, Joseph Kim, and Yulan He. 2022. PHEE: A dataset for pharmacovigilance event extraction from text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5571–5587, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ralph Sundberg and Erik Melander. 2013. Introducing the UCDP georeferenced event dataset. *Journal of Peace Research*, 50(4):523–532.

Beth M. Sundheim. 1992. Overview of the fourth Message Understanding Evaluation and Conference. In *Fourth Message Understanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*.

MeiHan Tong, Bin Xu, Shuai Wang, Meihuan Han, Yixin Cao, Jiangqi Zhu, Siyu Chen, Lei Hou, and Juanzi Li. 2022. DocEE: A large-scale and fine-grained benchmark for document-level event extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3970–3982, Seattle, United States. Association for Computational Linguistics.

Peter Turchin. 2012. Dynamics of political instability in the united states, 1780–2010. *Journal of Peace Research*, 49(4):577–591.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2022. DeepStruct: Pretraining of language models for structure prediction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 803–823, Dublin, Ireland. Association for Computational Linguistics.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, and 1 others. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Xinglin Xiao, Yijie Wang, Nan Xu, Yuqi Wang, Hanxuan Yang, Minzheng Wang, Yin Luo, Lei Wang, Wenji Mao, and Daniel Zeng. 2023. Yayi-uie: A chat-enhanced instruction tuning framework for universal information extraction. *arXiv preprint arXiv:2312.15548*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, and 1 others. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019. Doc2EDAG: An end-to-end document-level framework for Chinese financial event extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 337–346, Hong Kong, China. Association for Computational Linguistics.

A Details of Data Subset Selection

The details of some filters might be hard to understand without knowledge of UCDP design detailed in the codebook (Högbladh, 2023). Starting from the UCDP candidate events dataset, we apply the following filters:

- the UCDP annotator marked the event as “clear”,
- the event occurred on a single day (but this day might not be exactly known, thus the start date $\neq$ end date),
- the same event was not split into multiple entries (no “deaths split”),
- the source document uses the new post-2014 source format (ensuring more accurate publication dates),
- the event is described by a single source document (when using multiple sources, UCDP annotators might pick some of the information from one of the sources, distrust some specific roles from another, transforming the task into a weakly-supervised problem),
- the source document describes a single event (multi-event documents tend to include long tables of battle deaths without enough context to fully code the events),
- the start date, end date and publication date all occur in the same split,

- the deaths estimates are ordered properly: best (the sum of side A, side B, civilians and unknown) is in-between low and high,
- the split cut-off time are widened by 12 hours to avoid time zone reporting issues,
- the source document is in English,
- the length of the source document is between 100 and 10 000 characters.

All but the last two filters are applied on the UCDP database, resulting in 110 979 events. The final dataset is the result of HPLT substitution and the application of the last two filters on the HPLT source documents.

B Details of UCDP-AEC Fields

On domain openness UCDP-AEC is designed for a supervised prediction task where each instance consists of a source text and a publication date, denoted $\mathcal{X}$, and the corresponding output is a structured sequence of 14 argument roles $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_{14}$. Traditionally, when $\mathcal{Y}$ is a finite set, the problem is said to be closed domain, while potentially infinite $\mathcal{Y}$ is part of open-domain problems. In practice, open-domain roles may still be partially constrained because of real-world structures. For example, location values on the country level are often limited to internationally recognised countries.

The UCDP-AEC dataset contains fields with varying degrees of domain openness. To quantify the openness of each field in UCDP-AEC, we introduced three metrics in Section 4.3, report on each metric for each field in Table 2. We here describe these metrics in more detail.

We measure average instances per unique value, the *value density*, by computing the ratio $\frac{|\mathcal{D}|}{|\mathcal{Y}|}$, where $|\mathcal{D}|$ denotes the number of instances with a value for a given role, and $|\mathcal{Y}|$ the number of unique values observed. Higher values indicate a closed domain, with a small number of unique values that are frequently observed, whereas lower values indicate a more open domain with a more diverse set of values that are less frequently observed.

We measure the *unique value overlap* between the train and test set with the ratio $\frac{|\mathcal{Y}_{\text{test}} \cap \mathcal{Y}_{\text{train}}|}{|\mathcal{Y}_{\text{test}}|}$. This metric describes the percentage of test values observed during training, where a low overlap indicates a more open domain. A high overlap indicates that most of the values in the test set are also present in the train set.

Finally, for the *instance-level overlap*, we measure the proportion of instances in the test set with values, that has been observed in the train set with $\frac{|\mathcal{D}_{\text{test}} \cap \mathcal{D}_{\text{train}}|}{|\mathcal{D}_{\text{test}}|}$. Here, we take into account the frequency of the value.

In UCDP-AEC, the country-level location, and the death counts and estimates could be considered closed domain, as the range of possible values is limited and very repetitive.

The two actor roles, Side A and Side B, can be considered semi open domains, where many known actors will appear in the test set, but new actors will also occur. Both roles has a moderate value reuse, with about 26–27 instances per value. A similar pattern can be found for Subregion.

The more fine-grained locations show open domain characteristics. Region and Point are placenames often specific to single events, with many values in test that are not present in train. The temporal arguments have a 0% overlap because they are used for split selection (by definition, train and test do not overlap temporally).

On death estimates The fields “Deaths Low” and “Deaths High” are used when there is uncertainty with respect to the total number of fatalities. For these estimates, the detailed attribution of casualties to specific sides (Side A, Side B, Civilians, Unknown) is not specified. When summed together, those four deaths fields give the UCDP best estimate for the total number of people killed as part of an event. The fields “Deaths Low” and “Deaths High” are used when there is uncertainty with respect to the total number of fatalities. For these estimates, the detailed attribution of casualties to specific sides (Side A, Side B, Civilians, Unknown) is not specified. In any case we always have low $\leq$ best $\leq$ high and $1 \leq$ high.

C Sample Example

See Table 4. Note that we deliberately chose this article because of its short length for illustration purposes. Consisting of only 491 characters, this sample is not representative in terms of length, as the average article in UCDP-AEC contains 2116 characters.

D Detailed Results

See Table 5. Note that the accuracy for side A is given for the gold side A. Since actors are permutation invariants, this can correspond to a prediction for side B by the model. Side A and B are always different in the gold annotation, as such there is no

Input	
publication date	2022-01-06
source article	The lifeless body of a woman was discovered Wednesday morning on her farm in Mbengwi, Momo Division, Northwest region of Cameroon. Reports say there was a shootout between security forces and suspected separatist fighters around Womsei, a quarter behind the Mbon Market on Tuesday evening. Her body was only discovered Wednesday morning with the basket she was using to harvest okra still strapped on her back. Her body has been preserved at the Mbengwi mortuary pending burial arrangements.

Output			
Role	**Argument**	**Role**	**Argument**
Side A	Government of Cameroon	Start date	2022-01-04
Side B	Ambazonia insurgents	End date	2022-01-04
Location country	Cameroon	Location region	North-West region
Location subregion	Momo département	Location point	Mbengwi village
Deaths side A	0	Deaths side B	0
Deaths civilian	1	Deaths unknown	0
Deaths low	1	Deaths high	1

Table 4: Example of sample from the test set (sample id: 428220).

precision–recall trade-off to be taken advantage of despite evaluation through accuracy.

E Licences

The two data sources we are using are UCDP GED released under CC-BY and HPLT v2.0 realesed under CC0. We license our dataset UCDP-AEC under CC-BY.

We release our modifications to Text2Event and DEGREE under the same licence as the original code, that is MIT for Text2Event and Apache 2.0 for DEGREE.

We release the code of the Flan-T5 and Llama baselines, the code used to create the dataset, its evaluation script and scripts computing dataset statistics under the GNU AGPL licence.

F Evaluation Details

Since in UCDP-AEC all fields always have exactly one assigned value for all events, for a given field, the recall of a model is the number of true positive divided by the number of samples in the dataset, that is the accuracy. Similarly, the precision is also equal to the accuracy if the model predicts a single value for each field – as it should since its precision would only decrease otherwise – therefore we have precision $=$ recall $= F_1 =$ accuracy, thus explaining why we only report accuracies instead of the usual precision–recall–F_1 triplets.

For the date and deaths categories, we investigated the use of smooth metrics such as RMSE and RMSLE. However, as errors tend to be infrequent but important in magnitude, we did not find those metrics informative enough to be of interest. The provided evaluation script does compute RMSE and other metrics we considered such as semantic MRR.

For the semantic evaluation, we investigated using different models including S-BERT following Sharif et al. (2024). However we observed little variations apart from larger models providing slightly better P@1. Given the nature of the insight provided by the semantic evaluation, we did not find this warranted the extra computational cost and relied on `bert-base-uncased` instead.

G Model Details and Computational Budget

For both Text2Event (Lu et al., 2021) and DEGREE (Hsu et al., 2022), we reused the hyperparameters of the original models, except the input token limit, which we increased to 512 in both cases.

For the Flan-T5 baseline, we used learning rates of 3×10^{-5} with a maximum of 18 epochs. 10 intermediary evaluations were performed with early stopping on validation aggregate accuracy.

For the Llama-3.2-Instruct models, we set the maximum sequence length to 2048, used up to 900 training steps, applied a weight decay of 0.01, and a learning rate of 2e-4.

All models were trained on either P100 or A100 GPUs. Including development runs, the total GPU usage was around 200 GPU hours.

Model	Actors		Date		Location				Deaths					
	Side A	Side B	Start	End	Country	Region	Subregion	Point	Side A	Side B	Civilian	Unknown	Low	High
Text2Event	75.8	64.4	35.3	41.0	37.0	11.1	5.1	3.2	81.3	78.2	56.6	67.7	21.3	15.3
DEGREE	77.8	69.3	63.0	65.9	97.8	75.8	46.0	38.4	84.7	84.6	82.6	85.6	68.8	81.6
Flan-T5 large	77.0	70.8	60.3	65.7	97.8	73.3	45.7	34.3	82.3	81.9	82.9	87.3	69.8	82.1
Llama-3.2-I-1B	78.2	63.8	68.3	72.6	96.5	68.0	43.3	31.7	80.2	78.5	73.7	82.0	65.7	78.2
Llama-3.2-I-3B	78.9	67.1	72.8	74.7	97.1	76.7	47.5	34.4	83.9	80.3	75.5	83.0	68.6	79.7

Table 5: Accuracies for each individual role.

Instruction	Generate structured event information related to a socio-political conflict in the UCDP event format, based on the content of a source UCDP document. The generated data should include the following fields: `side_a_name`, `side_b_name`, `start_date`, `end_date`, `location_root_name`, `location_adm1_name`, `location_adm2_name`, `location_where_name`, `deaths_side_a`, `deaths_side_b`, `deaths_civilian`, `deaths_unknown`, `deaths_low`, `deaths_high`. Output the event information as a valid JSON object corresponding to the armed conflict incident described in the following source UCDP document.
Input	{Source article text}
Output	{Required output in JSON format}

Table 6: Instruction prompt used for fine-tuning.

H Instruction Prompt

The prompt used for the Llama models can be found in Table 6.

Americans are dreamers –
Generic statements and stereotyping in political tweets

Ines Rehbein
Data and Web Science Group
University of Mannheim
`rehbein@uni-mannheim.de`

Abstract

This paper addresses the question whether generics are associated with stereotyping, a claim often found in the literature. Most of the work, however, has been based on a small number of hand-crafted examples. We contribute to the discussion by looking at the use of generics in political tweets, written by members of the U.S. Congress, and present a framework for comparing the usage patterns of generics to other linguistic constructions that have been discussed as devices for stereotyping and *othering*. Our initial findings suggest that generics are not strongly associated with stereotypes while framing might play a more important role for stereotyping in political tweets.

1 Introduction

The construction of in-groups and out-groups, also refered to as *othering*, describes the categorisation of individuals as belonging to a specific group that is perceived as different from (and often inferior to) the social norm. In doing so, individuals considered as the *other* are usually depicted as a coherent and homogenuous group, based on stereotypes and overgeneralisation, making them an easy victim of discrimination (Staszak, 2009).

Othering is a common rhetorical move in political communication, and in particular in populist rhetoric (see, e.g., Fielder and Catalano (2017)). It can be strategically used to create a sense of shared identity among party supporters and mobilise them against a perceived enemy. While *othering* can have positive effects on the in-group, such as increasing group cohesiveness and cooperation, it also comes at the cost of prejudice, conflict and discrimination for the out-group.[1]

The creation of in- and out-groups plays an important role in political communication and is often associated with stereotyping. By *stereotype*, we refer to “a set of cognitive generalizations (e.g., beliefs, expectations) about the qualities and characteristics of the members of a group or social category”, following the definition of the APA Dictionary of Psychology.[2]

Several linguistic devices have been discussed as a means for othering and stereotyping, among them generic statements (Novoa et al., 2023; Bosse, 2024; Davani et al., 2024), the use of pronouns (Íñigo-Mora, 2004; Bull and Fetzer, 2006; Proctor and Su, 2011; Tyrkkö, 2016; Alavidze, 2017), framing by word choice (Sheshadri et al., 2021; van den Berg et al., 2020; Dreier et al., 2022), or entity framing (Entman, 1993; Mahmoud et al., 2025).

In the paper, we look at othering and stereotyping in political tweets written by members of the U.S. Congress. The focus of our work, however, is not so much on investigating stereotyping of in- and out-groups from a social science perspective but, instead, on studying the linguistic devices used for stereotyping in political tweets. Our work contributes to the linguistic discussion on whether generics are a common means to express stereotypes, as has often been argued (Geurts, 1985; Leslie, 2014; Radden, 2009; Novoa et al., 2023; Bosse, 2024; Ralston, 2024) but also refuted (Krifka et al., 1995).[3] In particular, we examine the role of generics, universals, and quantified statements based on real data from the political domain, where previous studies have mostly relied on a small number of constructed examples. We introduce a framework for extracting real-world generic statements about social groups and present a qualitative analysis of our findings.

---

[1] Early work investigating the effects of *othering* on human behaviour is the classic Robbers Case Study (Sherif et al., 1961) where young boys have been divided into groups to study how they behave when having to compete or cooperate with members of the out-group.

[2] `https://dictionary.apa.org/stereotype`.

[3] For a more detailed discussion, see Section 2.

Proceedings of the 5th Workshop on Computational Linguistics for the Political and Social Sciences (CPSS), pages 114–123
September 10-11, 2025 ©2025 Association for Computational Linguistics

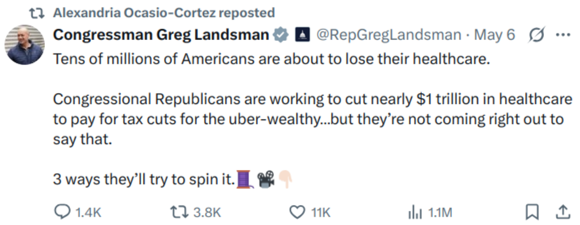
2 Related work

Generic statements are often described as a powerful tool to control the perception of groups as *in-groups* and *out-groups* by ascribing them certain characteristics and attributes and by creating an *otherness* (Staszak, 2009). One property that makes the use of generics so compelling is their tolerance toward exceptions. In contrast to universal statements like "all birds fly", which would be rejected by everybody with at least minimal knowledge of ornithology, the generic sentence "birds fly" will be accepted as true even though there are birds that do not have this capacity. Interestingly, this is also the case for statements that are true only for a small subset of its group members, such as "mosquitoes transfer the West Nile Virus" which holds for only a small percentage of mosquitoes.

In addition, Novoa et al. (2023) point out another striking feature of generics, namely that "despite expressing general claims that gloss over exceptions, they are endorsed in the face of variable or even minimal evidence". This tolerance to exceptions and ignorance towards a lack of evidence seems to make generics especially suited to transport stereotypes. Bosse (2024, p.3878), for example, describes generics as "mental states that associate properties with social groups".[4] However, there are also other means to associate properties with groups, such as explicitly quantified statements ("most birds can fly") and universals ("all birds can fly").

Othering and generics have both been discussed in relation to polarisation (Erdoğan and Uyan-Semerci, 2025; Iyengar and Westwood, 2015). Novoa et al. (2023) investigate the role of generics for partisan polarisation in the US and show that not only generics but also explicitly quantified statements can be interpreted by humans as universally true, pointing at a tendency of the human brain to convert more nuanced, quantified statements into generic, categorising ones. In three experiments, the authors show that generic expressions and, to a lesser extent, explicitly quantified statements can evoke polarised political judgments.

Some works in NLP have focussed on the detection of stereotypes (Blodgett et al., 2020; Sap et al., 2020), often using artificial data based on *(identity term, attribute)* tuples (Jha et al., 2023). One recent example is the Multilingual SeeGULL dataset of Bhutani et al. (2024) which includes LLM generated tuples, validated by human coders. Davani et al. (2024) focus on the use of generic sentences for stereotyping and present a multilingual dataset that combines specific group mentions with attributes, annotated for genericity. Their schema also distinguishes mentions of generalisations (e.g., reporting what others have said) from promoting it. The authors show that the co-occurrence of group mentions and attributes are not a reliable incdicator of generalisation, thus questioning the widespread approach of using group, attribute pairs for stereotype detection. They present a classifier that detects generalisations, reporting a PR-AUC (Area Under the Precision Recall (PR) Curve) of 58.7.


Alexandria Ocasio-Cortez reposted

Congressman Greg Landsman @RepGregLandsman · May 6

Tens of millions of Americans are about to lose their healthcare.

Congressional Republicans are working to cut nearly $1 trillion in healthcare to pay for tax cuts for the uber-wealthy...but they're not coming right out to say that.

3 ways they'll try to spin it.🧵📽️👇

Figure 1: Example tweet including a numerically modified statement about Americans.

More linguistically inspired work has focussed on the annotation and detection of generics in language (Mitchell et al., 2003; Doddington et al., 2004; Walker et al., 2006; Suh, 2006; Herbelot and Copestake, 2008; Zirn et al., 2008; Mathew and Katz, 2009; Reiter and Frank, 2010; Friedrich et al., 2015b; Friedrich and Pinkal, 2015; Friedrich et al., 2016). We take this as the starting point for our framework to extract generic and generalising statements from real-world political text.

3 Framework

We now describe the different steps in our framework for extracting instances of generic, universal and quantified statements about groups, in order to investigate their function in discourse (Figure 2).

Data The data we use in our case study are daily tweets of both houses of congress, from over 1,000 campaign, office, committee and party accounts.[5] The data has been collected over a time period from June 2017 to October 2020 and amounts to ca. 9.8 mio tweets. For an example, see Figure 1.

---

[4] Others, however, have questioned the idea that generics express stereotypes. See, e.g., Krifka et al. (1995, p.48).

[5] The data is available from `https://github.com/alexlitel/congresstweets`.

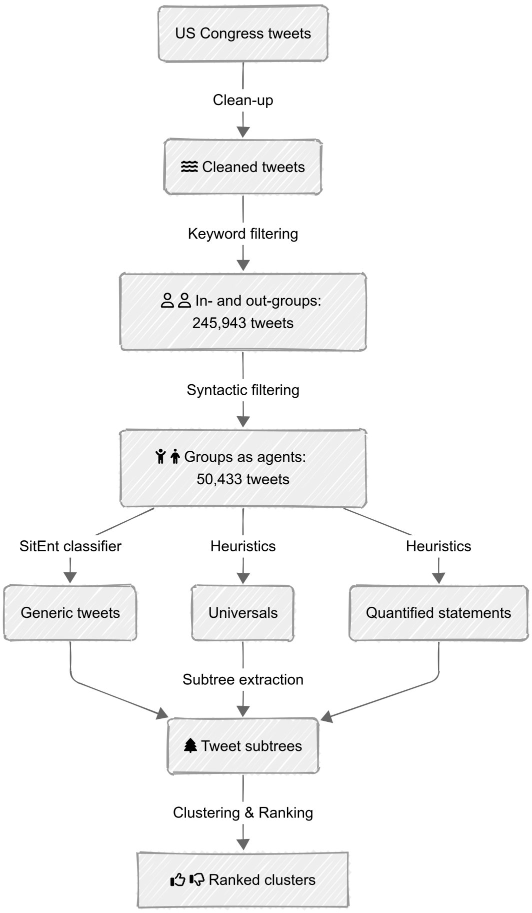


Figure 2: Framework for the extraction of real-world linguistic examples from U.S. Congress tweets.

Preprocessing We do some basic clean-up during preprocessing where we extract the tweet messages and filter out noise. Preprocessing includes the removal of non-ascii characters, empty lines and punctuation at the beginning of a tweet. We remove duplicates and tweets with no lowercase letters and delete URLs, at-mentions and hashtags at the beginning or end of the tweet.

Keyword-based filtering In the next step, we collect a list of keywords representing in- and out-groups in the migration discourse, i.e., *alien, american, asylum, migrant, refugee*. We then do a string-based keyword search for these group terms, ignoring case, and extract all tweets that include (at least) one of the terms. Please note that this also extracts tweets that include collocations like *American dream, American economy* or *immigrant rights* which are not in the center of our interest. We remove those in the next step, syntactic filtering. Keyword-based filtering reduces the overall amount of tweets from 9.8 mio to a much smaller set of 245,943 tweets (see Table 1) that can be processed more efficiently.

	group	is-subj	generic
alien	2,593	224	92
american	205,986	46,500	8,525
asylum	5,781	362	166
migrant	26,996	2,895	1,565
refugee	4,587	452	214
total	245,943	50,433	10,562

Table 1: Number of tweets after the different filtering steps (group: keyword-based filtering; is-subj: tweets where the group term is in subject position; generic: generic and generalising statements, according to the predicted Situation Entities classifier).

Syntactic filtering Next, we use a dependency parser[6] to extract tweets where the group terms occur in subject position. This filters out adjectival uses of the terms and extracts plausible candidates for generic statements about groups. We specify two exceptions, *American people, asylum seeker*, that we include in the data. This syntactic filtering step further reduces the data from roughly 246,000 messages to 50,433 tweets where one of the group terms appears in subject position in the tweet.

3.1 Classifying genericity

To identify generic and generalising statements, we train a supervised text classifier on the Situation Entities (SitEnt) dataset (Friedrich and Palmer, 2014; Friedrich et al., 2015b). Situation Entities are clause-level semantic annotations that introduce *situations* to the discourse (Smith, 2003), such as EVENTS, STATES, GENERICS, GENERALIZING SENTENCES, REPORTS or QUESTIONS (see A.1). We use a simple text classifier, based on a pre-trained RoBERTa model (Liu et al., 2019). For more training details, see §A.2 in the appendix.

Table 2 shows results on the SitEnt test set for all classes. Overall, results are high with an accuracy of 0.8 and a weighted F1 of 0.79. Not surprisingly, we get higher scores for classes with more training instances and also high results for Situation Entity types that have some easily identifiable features (such as questions, reports or speech acts). Results for GENERIC are substantially lower than for EVENT, even though both classes have roughly the same number of instances, and generalising sentences seem even harder to detect.

Our results are in the same range as the state-of-the-art (Rezaee et al., 2021) which combines a BERT model (Devlin et al., 2019) with GPT-2 (Radford et al., 2019) and reports an F1 of 0.79

---

[6] We use the spaCy v2.3.9 dependency parser with the English model `en_core_web_sm`.

	Prec	Rec	F1	Support
STATE	0.82	0.89	0.85	3,174
EVENT	0.85	0.83	0.84	1,524
GENERIC	0.80	0.72	0.76	1,429
UNDECIDED	0.67	0.63	0.65	465
GENERALISING	0.54	0.47	0.50	294
REPORT	0.86	0.81	0.84	274
QUESTION	0.89	0.92	0.91	132
GENERALSTATIVE	0.00	0.00	0.00	86
IMPERATIVE	0.51	0.90	0.65	51
SPEECHACT	1.00	1.00	1.00	1
Accuracy			0.80	7,430
F1(weighted)	0.79	0.80	0.79	7,430

Table 2: Results for the RoBERTa-based Situation Entity classifier on the test set. Support shows the number of instances for each class in the test set.

(acc: 0.81) on the SitEnt corpus.[7] The lower results for generic and generalising statements shows the difficulty of the task, which is in line with previous work (e.g., Davani et al. (2024), see §2 above). For an error analysis, see §A.3.

We apply our classifier to the Twitter dataset and extract all messages that have been predicted as either generic or generalising. As a result, we obtain a dataset with 10,562 tweets (see Table 1).

3.2 Extracting clusters of similar sentences

To find out whether there is an association between form and function for the different linguistic devices (i.e., generic, universal and quantified statements), we need to compare those constructions in context, i.e., in sentences with similar meaning. To extract those, we proceed as follows.

Subtree extraction As input, we use the set of tweets where instances of each group term appear in subject position. This includes all different statement types. Next, we identify the verb governing the group term and extract its syntactic subtree. This can be done, using the spaCy subtree function which extracts a token and all its syntactic dependents. For the example sentence below, the extracted subtree is the clause that includes the group term in subject position (Ex 1).

Ex 1. "Climate change is expected to create hundreds of millions of refugees, but right now climate refugees have no legal rights."
Subtree: "refugees have no legal rights"

Clustering We encode the subtree strings, using sentence embeddings, and cluster the embeddings to obtain semantically similar sentences.[8] For additional details on the clustering process, see A.4. We use a high cosine similarity threshold for clustering. As a result, we loose many instances that are below the threshold. However, for our use case, we are not interested in high coverage (i.e., being able to assign every instance to a cluster), but rather in identifying usage patterns of similar sentences in order to examine the variation in the use of different constructions. Table 3 shows an example cluster of statements expressing the americans' support for law and order. All but one sentence in this cluster use generic statements while the remaining one uses a quantified statement.

Ranking Finally, we rank the resulting clusters to identify similar sentences that show a high preference for specific constructions (i.e., a strong *form-to-meaning association*), providing us with linguistic evidence for further qualitative analysis. For ranking, we use the ratio of instances for the respective construction (i.e. generic, universal, etc.), relative to the total number of instances in the cluster.[9] To identify generic and generalising statements, we rely on the predictions of the SitEnt classifier (see §3.1). Universal and quantified statements can be easily extracted, based on heuristics. For universal statements, we extract sentences where the group term in subject position is modified by either *all* or *every* while for quantification, we search for the modifiers *most, many, some, few, several*.

4 Exploration of generics and stereotypes in political tweets

To explore to what extend generics are used to stereotype social groups in political tweets, we apply our framework to the U.S. Congress Tweets dataset described above. We present a qualitative analysis, focussing on clusters of similar sentences that show a strong preference for one particular construction.

---

[7] Please note that we do not report averaged results over different model initialisations, as we are not interested in benchmarking our models on the Friedrich et al. (2015b) data but only in identifying and extracting generic and generalising statements from social media.

[8] We use the Fast Clustering algorithm `https://sbert.net/examples/sentence_transformer/applications/clustering/README.html#fast-clustering` from the Sentence Transformer (Reimers and Gurevych, 2019) library.

[9] There are many more sophisticated measures that could be used, however, for the purpose of exploration our simple ranking metric already works well in identifying semantic clusters with a strong preference for one particular construction.

ID	**Cluster 1112**, 10 elements ratio: 0.9
1	Americans want LAW & ORDER!
2	The American people stand with law enforcement.
3	Americans support law enforcement and want law & order.
4	Americans support our # police officers and law and order in our cities.
	...
10	**Most Americans** support law & order, the fair application of it, and pray for our police.

Table 3: Example cluster for the group term "american", including 9 generic sentences (ratio: 0.9) and one explicitly quantified statement (**Most** Americans).

Cluster 267, 23 elements ratio: 1.0
ALL Americans have the constitutional right to marry
ALL Americans have the right to marry no matter whom they love
Cluster 813, 11 elements ratio: 1.0
all Americans have the opportunity to achieve success
every American has the opportunity to succeed
Cluster 838, 11 elements ratio: 1.0
all Americans can enjoy freedom & equality
every American is equal

Table 4: Examples for clusters with a high ratio of universal statements.

Generics Generic statements about the in-group in our data mostly express political demands (see A_{1-5} below).

Americans

(A_1) deserve better

(A_2) want the truth / want transparency

(A_3) should know / need to know

(A_4) are tired of waiting

(A_5) need help

For the out-group, clusters that show a strong preference for generics include mostly positive statements about the group (see B_{1-4} below).

Immigrants / refugees / asylum seekers

(B_1) enhance our economy

(B_2) make America stronger

(B_3) are not criminals / animals

(B_4) deserve protection

We did not find instances of generics used for negative stereotyping of social groups.

Universals Table 4 shows examples for semantic clusters that exclusively included universals. It is obvious that those instances do not contain stereotypes but express normative generalisations about how the world *should be*, rather than how it is. All of the clusters with a high ratio of universal sentences in our analysis are of that type.

Aliens	224 mentions in subject position:
	illegal (155), criminal (30), deported (4), convicted (3), dangerous (3), charged (2)
Immigrants	2,917 mentions in subject position:
	illegal (223), undocumented (154), young (62), legal (30), working (14)
Asylum s.	365 mentions in subject position:
	vulnerable (5), fleeing (5), cuban (3), transgender (3), legitimate (3)
Refugees	457 mentions in subject position:
	welcome (15), fleeing (14), rohingya (9), syrian (9), american (4), somali (4), liberian (4), clean (3)

Table 5: Entity framing of social groups in tweets.

Quantified statements When comparing generics to sentences that include explicit quantification, we did not see similarly strong associations between form and meaning (also see Table 9 in the apendix).

Entity framing While we did not see strong evidence for an assiciation between generics and stereotypes in our data, we found numerous instances of entity framing. Table 5 shows the most frequent modifiers for each group term, extracted from the spaCy dependency parse. It is obvious that the terms *aliens, immigrants* are mostly framed in negative terms. *Aliens* in particular are portrayed as dangerous and criminal, which is reflected in framing through word choice. Politicians with a more positive stance towards immigration did not use the term *alien* but preferred *refugee*, shown by the lack of negative attributes for this term.

5 Conclusion

In the paper, we present a framework for extracting generic and generalising statements about social groups and apply it to tweets written by members of the U.S. Congress. Our framework allows us to extract linguistic evidence for the various constructions while controlling for sentence meaning. This enables us to study usage patterns for generics, universals and quantified statements in a controlled setting with real-world data. While generics are often considered as an important device for stereotyping, our initial findings suggest that the association is less strong as expected and that framing might play a more important role for stereotyping social groups in political tweets.

Limitations

We would like to point out some limitations of our work. First, we would like to emphasize once again that the framework is not suitable for comparing the use of stereotypes across parties or across time. In the paper, we only consider three constructions that have the potential to be used for stereotyping but do not aim to measure the total amount of stereotypical statements overall.

Second, while this short paper introduces a framework for extracting real-world examples in context for comparative linguistic analysis, it does not (yet) present a validation of the framework. Third, while the clustering step manages to produce clusters of sentences with similar meaning, it is conceivable that the surface form of the different constructions also has some impact on the clustering process and that this results in artificially "clean" groups of statements where the different constructions end up in separate clusters. One way to address this issue is to mask the markers for the various constructions so that the input to the clustering process is normalised (see example below) while keeping the information on the original statement type for further analysis.

Orig: *All|Most|∅* Americans support the law.
Cluster input → Americans support the law.

Acknowledgments

The work presented in this paper is funded by the German Research Foundation (DFG) under the UNCOVER project (RE3536/3-1). We would also like to thank the anonymous reviewers for their constructive feedback.

References

Maia Alavidze. 2017. The use of pronouns in political discourse. *International Journal of Arts & Sciences*, 9(4):349–356.

Mukul Bhutani, Kevin Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. 2024. SeeGULL multilingual: a dataset of geo-culturally situated stereotypes. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 842–854, Bangkok, Thailand. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Anne Bosse. 2024. Stereotyping and generics. *Inquiry: An Interdisciplinary Journal of Philosophy*, 67(10):3876–3892.

Peter Bull and Anita Fetzer. 2006. Who are we and who are you? the strategic use of forms of address in political interviews. *Text and Talk*, 26.

Aida Mostafazadeh Davani, Sagar Gubbi, Sunipa Dev, Shachi Dave, and Vinodkumar Prabhakaran. 2024. Genil: A multilingual dataset on generalizing language. *Preprint*, arXiv:2404.05866.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Sarah K Dreier, Emily K Gade Dallas Card, and Noah Smith. 2022. Patterns of bias: How mainstream media operationalize links between mass shootings and terrorism. *Political Communication*.

Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4):51–58.

Emre Erdoğan and Pinar Uyan-Semerci. 2025. Othering in politics: how affective polarization undermines democratic philia? *Frontiers in Political Science*, pages 1–17.

Grace E. Fielder and Theresa Catalano. 2017. Othering others: Right-wing populism in uk media discourse on "new" immigration. In Jan Chovanec and Katarzyna Molek-Kozakowska, editors, *Representing the Other in European Media Discourses*, pages 207–234. Amsterdam: John Benjamins.

Annemarie Friedrich, Kleio-Isidora Mavridou, and Alexis Palmer. 2015a. Situation entity types (annotation manual). Version 1.1.

Annemarie Friedrich and Alexis Palmer. 2014. Situation entity annotation. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 149–158, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Annemarie Friedrich, Alexis Palmer, Melissa Peate Sørensen, and Manfred Pinkal. 2015b. Annotating genericity: a survey, a scheme, and a corpus. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 21–30, Denver, Colorado, USA. Association for Computational Linguistics.

Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. 2016. Situation entity types: automatic classification of clause-level aspect. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1757–1768, Berlin, Germany. Association for Computational Linguistics.

Annemarie Friedrich and Manfred Pinkal. 2015. Discourse-sensitive automatic identification of generic expressions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1272–1281, Beijing, China. Association for Computational Linguistics.

Barts Geurts. 1985. Generics. *Journal of Semantics*, 3(4):247–255.

Aurelie Herbelot and Ann Copestake. 2008. Annotating genericity: How do humans decide? (a case study in ontology extraction). In Sam Featherston and Susanne Winkler, editors, *The Fruits of Empirical Linguistics*, volume 1. Berlin: de Gruyter.

Isabel Íñigo-Mora. 2004. On the use of the personal pronoun we in communities. *Journal of Language and Politics*, 3(1):27–52.

S Iyengar and S Westwood. 2015. Fear and loathing across party lines: new evidence on group polarization. *American Journal of Political Science*, 59:690–707.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870, Toronto, Canada. Association for Computational Linguistics.

Manfred Krifka, Francis J Pelletier, Gregory N Carlson, Gennaro Chierchia, Godehard Link, and Alice ter Meulen. 1995. Genericity: An introduction. In *The Generic Book*, pages 1–24. University of Chicago Press, Chicago and London.

Sarah-Jane Leslie. 2014. Carving up the social world with generics. *Oxford studies in experimental philosophy*, 1:208–232.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Tarek Mahmoud, Zhuohan Xie, Dimitar Dimitrov, Nikolaos Nikolaidis, Purificação Silvano, Roman Yangarber, Shivam Sharma, Elisa Sartori, Nicolas Stefanovitch, Giovanni Da San Martino, Jakub Piskorski, and Preslav Nakov. 2025. Entity framing and role portrayal in the news. *Preprint*, arXiv:2502.14718.

Thomas A. Mathew and Graham Katz. 2009. Supervised categorization for habitual versus episodic sentences. In *Sixth Midwest Computational Linguistics Colloquium*. Bloomington, Indiana: Indiana University.

Alexis Mitchell, Stephanie Strassel, Mark Przybocki, JK Davis, George R. Doddington, Ralph Grishman, Adam Meyers, Ada Brunstein, Lisa Ferro, and Beth Sundheim. 2003. Ace-2 version 1.0. LDC2003T11.

Gustavo Novoa, Margaret Echelbarger, Andrew Gelman, and Susan A. Gelman. 2023. Generically partisan: Polarization in political communication. *Proceedings of the National Academy of Sciences*, 120(47):e2309361120.

Katarzyna Proctor and Lily I-Wen Su. 2011. The 1st person plural in political discourse – American politicians in interviews and in a debate. *Journal of Pragmatics*, 43(13):3251–3266.

Günter Radden. 2009. Generic reference in english: A metonymic and conceptual blending analysis. In Klaus-Uwe Panther, Linda L. Thornburg, and Antonio Barcelona, editors, *Metonymy and Metaphor in Grammar*, pages 199–228. John Benjamins Publishing Company.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*.

Tom Ralston. 2024. Reconceptualising the psychological theory of generics. *Philosophical Studies*, 181:2973–2995.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reiter and Anette Frank. 2010. Identifying generic noun phrases. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 40–49, Uppsala, Sweden. Association for Computational Linguistics.

Mehdi Rezaee, Kasra Darvish, Gaoussou Youssouf Kebe, and Francis Ferraro. 2021. Discriminative and generative transformer-based models for situation entity classification. *CoRR*, abs/2109.07434.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Muzafer Sherif, O.J. Harvey, Jack B. White, William R. Hood, and Carolyn W. Sherif. 1961. *Intergroup Conflict and Cooperation. The Robbers Cave Experiment [1954]*. University of Oklahoma Book Exchange, Norman.

Karthik Sheshadri, Chaitanya Shivade, and Munindar P. Singh. 2021. Detecting framing changes in topical news. *IEEE Transactions on Computational Social Systems*, 8(3):780–791.

Carlota S Smith. 2003. *Modes of discourse: The local structure of texts*. Cambridge University Press.

Jean-François Staszak. 2009. Other/otherness. In Rob Kitchin and Nigel Thrift, editors, *International Encyclopedia of Human Geography*, pages 43–47. Elsevier, Oxford.

Sangweon Suh. 2006. Extracting generic statements for the semantic web. Master's thesis, University of Edinburgh. Master's thesis.

Jukka Tyrkkö. 2016. Looking for rhetorical thresholds: Pronoun frequencies in political speeches. *Studies in Variation, Contacts and Change in English*, 17.

Esther van den Berg, Katharina Korfhage, Josef Ruppenhofer, Michael Wiegand, and Katja Markert. 2020. Doctor who? framing through names and titles in German. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4924–4932, Marseille, France. European Language Resources Association.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. LDC2006T06.

Cäcilia Zirn, Vivi Nastase, and Michael Strube. 2008. Distinguishing between instances and classes in the wikipedia taxonomy. In *Proceedings of the 5th European Semantic Web Conference on The Semantic Web: Research and Applications*, ESWC'08, page 376–387, Berlin, Heidelberg. Springer-Verlag.

A Appendix

A.1 Situation Entity types

Table 6 illustrates the different Situation Entity types. Most relevant for our work are the classes GENERIC and GENERALISING. GENERIC sentences have a generic main referent while GENERALISING sentences make statements about individuals that generalise over situations and report regularities.

A.2 Classification setup

We follow the suggested train-test splits,[10] resulting in 30,859 instances for training and 7,430 instances for testing. We further split the training data into a training and development set (29,316 and 1,543 instances).

Preprocessing The original dataset has been automatically segmented into clauses. As the segmentation step introduces noise, we do not rely on the provided clause boundaries but extract the left and right context within the sentence for each clause and mark the relevant clause information by inserting an underscore before the verbal head of the clause. This means that the same sentence can be included multiple times, but with different verbs highlighted within the sentence. This enables the model to learn which clause to focus on while, at the same time, providing the necessary contextual information.

Hyperparameter tuning We do hyperparameter tuning on the development set and obtain best results of 0.86 F1(micro) and 0.63 (macro) across all 10 labels. We use the following hyperparameter configuration:

- learning rate: 3.134214402336582e-05
- weight decay: 3.215675091397119e-05
- batch size: 16
- warmup ration: 0.0936967038422192
- number of train epochs: 2

For implementation, we use the huggingface transformers library: `https://huggingface.co/docs/transformers`.

A.3 Error analysis

The training data for the Situation Entity classifier includes texts from 13 different genres and text types, including blogs, emails, essays, fiction, letters, news, and wikipedia articles. To get an idea about the error rate for predicting generics and generalising sentences, we randomly sample 20 instances for each group and manually inspect the classifier's predictions.

Our error analysis shows that the overall accuracy of the classifier on the twitter data is quite high with 0.84. For the term *aliens*, most instances are predicted as GENERALISING as this term is mostly

---

[10] `https://github.com/annefried/sitent/tree/master`.

SitEnt type	Definition	Example
STATE	States introduce specific properties of specific individuals to the discourse.	Joseph loves cake.
EVENT	Events introduce a specific event to the discourse: things that happen or happened.	Mira won the race.
GENERIC	Generic sentences make statements about kinds.	Lions are carnivores.
GENERALISING	Generalising sentences report regularities related to specific individuals.	Rob often feeds my cat.
REPORT	SitEnt introduced by verbs of speech.	"Let's go home" said Tim.
SPEECH ACTS		
IMPERATIVE	Commands, requests, advice.	Stay calm.
QUESTION	Functional speech act to elicit information.	Are you tired?

Table 6: Description and examples for the different Situation Entity types, adapted from Friedrich et al. (2015a) (available from `https://www.coli.uni-saarland.de/projects/sitent/page.php?id=resources`).

group term	correct	acc.
alien	15	0.75
american	18	0.90
asylum	16	0.80
migrant	18	0.90
refugee	17	0.85
all groups	84	0.84

Table 7: Number of correctly predicted labels (predicted as either GENERIC or GENERALISING by the classifier) for a random sample of 100 instances (20 instances per group) and accuracy for each social group.

used in combination with the adjective *illegal*, thus making this a reference not to a kind but to a specific subgroup of the kind (here: all aliens that are illegal). This is different from the other group terms where the main referent is generic at least for half of the instances (*Americans*) or for the majority of the messages (*asylum seekers, migrants, refugees*).

We observe the highest error rate for *asylum seekers* where our filtering heuristic does not seem to be accurate enough. Here we find 4 instances where the term *asylum* does not refer to the group, thus making the extracted instances irrelevant. Some other errors are due to mistakes made by the dependency parser failing to identify the correct subject.

A.4 Clustering similar clauses

We use the Fast Clustering algorithm[11] from the Sentence Transformer (Reimers and Gurevych, 2019) library. Cluster similarity is determined, based on cosine similarity. We set the minimum cluster size to 10 and the similarity threshold to 0.8, meaning that sentences need to have a similarity of ≥ 0.8 to be considered as similar. Sentences that do not meet this criterium are ignored.

group term	no. clusters	no. instances
alien	0	0
american	1,129	29,027
asylum	4	215
migrant	53	1,402
refugee	8	229

Table 8: Number of clusters and clustered instances for each social group term.

We cluster the tweets for each group term separately and use the `all-MiniLM-L6-v2` model[12] to map the extracted clauses to a 384 dimensional dense vector space, as input for clustering. Table 8 shows the number of clusters and clustered instances for each group.

For the term *alien*, we do not get any clusters, due to the small number of instances and rather high similarity threshold. However, for our use case we are not interested in high coverage (i.e., being able to assign each instance to a cluster) but rather in identifying usage patterns of similar sentences to study the variation in the use of the different constructions.

Table 9 shows examples for the clusters with the highest ratio of quantified statements for the term *American*. We only find two clusters where the ratio of quantified statements is higher than 0.5, indicating that explicit quantification does not seem to have a similar function in the discourse as universals.

---

[11] `https://sbert.net/examples/sentence_transformer/applications/clustering/README.html#fast-clustering`.

[12] `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

Cluster 551, 14 elements	ratio: 0.86
Americans are struggling to pay their rent and mortgages.	
Many Americans are struggling to pay their rent or mortgage mortgage.	
Cluster 725, 12 elements	ratio: 0.73
The American people overwhelmingly support this measure .	
most Americans from both parties are in support	
Cluster 1072, 10 elements	ratio: 0.30
Americans still face too many barriers at the ballot box	
too many Americans still face barriers at the ballot box with voter ID	

Table 9: Examples for clusters with a high ratio of quantified statements.

Author Index

Arefyev, Nikolay, 98

Bechara, Hannah Dorothy, 30
Braun, Laura, 23

Carreño, Ramón, 98

Dakota, Daniel, 1

Garcia-Corral, Paulina, 30

Heid, Ulrich, 11

Jankin, Slava, 30

Knierim, Aenne Cecilia Kristine, 11, 11
Kübler, Sandra, 1

Mandl, Thomas, 11
Mandravickaitė, Justina, 80
Manohara, Krishnamoorthy, 30
McCrae, John Philip, 68
Miehling, Daniel, 1
Mishra, Rahul, 98

Olsen, Helene Bøsei, 98
Oswald, Christian, 23
Øvrelid, Lilja, 98

Rehbein, Ines, 114
Rilla, Raluca, 48

Simon, Étienne, 98
Stede, Manfred, 48
Steinbach, Hannah Mathilde, 48

Velldal, Erik, 98

Waskow, M.a., 68

Yilmaz, Mert Can, 98
Yüzüncüoglu, Imge, 48