Leveraging Secure Social Media Crowdsourcing for Gathering Firsthand Account in Conflict Zones

Abanisenioluwa Orojo[1] [0009−0005−9448−1929], Pranish Bhagat[1], John Wilburn[2], Michael Donahoo[1], and Nishant Vishwamitra[2]

[1] {abanisenioluwa_oroj1, pranish_bhagat1, jeff_donahoo}@baylor.edu
Baylor University, United States
[2] nishant.vishwamitra@utsa.edu, john.wilburn@my.utsa.edu
University of Texas at San Antonio, United States

Abstract. The Russo-Ukrainian conflict underscores challenges in obtaining reliable firsthand accounts. Traditional methods such as satellite imagery and journalism fall short due to limited access to zones. Secure social media platforms such as Telegram offer safer communication from conflict zones but lack effective message grouping, hindering insight collection. The proposed framework aims to enhance firsthand account gathering by crowdsourcing secure social media data. We gathered 250,000 Telegram messages on the conflict and developed a language model-based framework to identify contextual groupings. Evaluation reveals 477 new groupings from 13 news sources, enriching firsthand information. This research emphasizes the significance of secure social media crowdsourcing in conflict zones, paving the way for future advancements.

Keywords: Crowdsourcing · Social Media · Conflict Zones · Firsthand Accounts · Secure Communication · Data Analysis · Natural Language Processing.

1 Introduction

The intricate fabric of today's global societal landscape is increasingly dominated by geopolitical turmoil that dramatically affects societies on multiple levels. In this age of rapidly escalating conflicts, exemplified by situations such as the 2022 Russian invasion of Ukraine [6], gathering precise firsthand accounts from these volatile areas has never been more critical. As complexities grow, the need for nuanced, timely insight into conflict zones increases.Traditionally, firsthand intelligence gathering relied on satellite imagery, on-the-ground journalism, and diplomatic sources [1]. These methods offer a limited lens to view unfolding realities, as turbulent situations in conflict zones prohibit public surveying. They often provide a delayed, fragmented picture, lacking the full spectrum of human experiences within conflict zones. The digital age has ushered in new data sources to enrich understanding of conflict-ridden regions [14]. Secure social media like

Telegram has emerged as a powerful tool [12], offering people to safely communicate first-hand accounts, resulting in a stream of real-time, user-generated content reflecting diverse viewpoints. Studies show these platforms are increasingly used in conflict zones to communicate vital firsthand information. Yet, with vast data comes a challenge: extracting meaningful insights from complex data streams [4]. A major challenge to extracting meaningful insights from secure social media datastreams is the lack of organized communication patterns in groups where such information from the ground is shared. Relevant firsthand account reports are often lost among other conversations. Techniques that can capture these multiple interweaving conversations are needed to extract meaningful interactions and key themes [13]. We propose a framework to crowdsource firsthand account information from secure social media in conflict zones. Our framework uses TF-IDF [10] to extract key phrases about a conflict. It then organizes messages using a novel conversations extraction algorithm based on RoBERTa [8], identifying firsthand reports and separating them from irrelevant conversations. We address the following key research questions: **RQ1:** How can secure social media communication from conflict zones be organized to effectively augment existing open-source information? **RQ2:** Can the integration of crowdsourced information from secure social media applications enhance the richness and comprehensiveness of open-source data? We focus on the Russia-Ukraine conflict[3] and collect a dataset of 250,000 Telegram (*i.e.*, a social media platform popularly used for secure communication) messages[4]. Our framework addresses RQ1 by identifying conversations pertaining to firsthand information, outperforming baselines by 158.13%. For RQ2, we capture 477 new conversational groups with key insights on specific events, such as bombing of Ukrainian cultural heritage sites and health facilities. [5]

2 Related Works

Our exploration encompasses four domains within conflict research: social media's role, conflict data collection methodologies, empirical data use, and human rights violations detection via social media. Social media significantly influences modern conflicts, serving as a critical communication tool. It impacts conflicts by lowering communication costs, accelerating information dissemination, prompting strategic adaptation, and offering new data [15]. However, it also fuels conflict through disinformation and extremist recruitment. Peacebuilding efforts are incorporating digital perspectives to counteract these challenges.

Key strategies prioritize accuracy, consistency, and replicability [13]. The inclusion of "non-events" and intercoder reliability are crucial, especially for social media data interpretation. High-quality, disaggregated data is fundamental for

[3] Our framework can be generically applied to any conflict. We use the Russia-Ukraine conflict to demonstrate our approach.

[4] Our dataset will be made publicly available.

[5] Complete Figures, Tables, and additional References can be found at: https://github.com/AKOrojo/ASONAM2024-Secure-Crowdsourcing.git

understanding conflict processes [5]. The advent of ’big data’ has necessitated rethinking data aggregation methods. Challenges in collecting high-quality conflict data persist due to observation difficulties and varied information sources. Social media platforms have proven effective in uncovering evidence of atrocities in conflict zones. However, challenges include information overload, data reliability, verification, admissibility in legal contexts, and potential underreporting due to technology access disparities.

Table 1. Samples of Articles and Discussed Topics

Name	**Topics Discussed**	**Date**
Ukrainian Cultural Heritage Potential Impact Summary	impacts on Ukrainian cultural heritage, climate and gastronomy.	May 2022
Kyiv Falling into Darkness	Instability and decreased light production in Kyiv.	Nov 2023

3 Methodology

3.1 Data Collection

Our data collection process involves three streams: satellite articles, Telegram messages, and a social media conversations dataset. The Yale Humanitarian Research Lab’s Conflict Observatory provided satellite articles [7] covering the Russia-Ukraine conflict from May 2022 to 2023 (Table 1). Telegram messages were collected via Lyzem [9], totaling 258,101 messages from 2020 to Aug 2023, with 150,083 in 2022 and 107,351 in 2023. We also utilized Cornell University’s ConvoKit Social Media Conversations Dataset [2] to train our LM algorithm, creating a dataset of message pairs labeled for conversational context. This comprehensive approach combines traditional reporting methods with secure social media data to provide a holistic view of the conflict.

3.2 Overview of Our Approach

Our framework comprises five main components: (1) Data Acquisition, which ingests Yale Conflict Observatory articles and Telegram messages; (2) Preprocessing & Keyword Extraction, involving data cleaning and TF-IDF application for keyword extraction; (3) Message Searching, using extracted keywords to identify relevant Telegram messages; (4) Conversation Modeling & Thread Identification, employing an LM-based model to identify conversation threads in message clusters; and (5) Topic Analysis, performing analysis on identified threads to extract firsthand accounts and key discussion topics. This approach integrates traditional conflict reporting with social media data analysis to provide a comprehensive view of the conflict landscape.

3.3 Preprocessing & Keyword Extraction

The initial stages involve collecting, cleansing, and structuring data to facilitate further analysis. This process prepares the data by removing extraneous information and organizing it efficiently. The keyword extraction was performed on

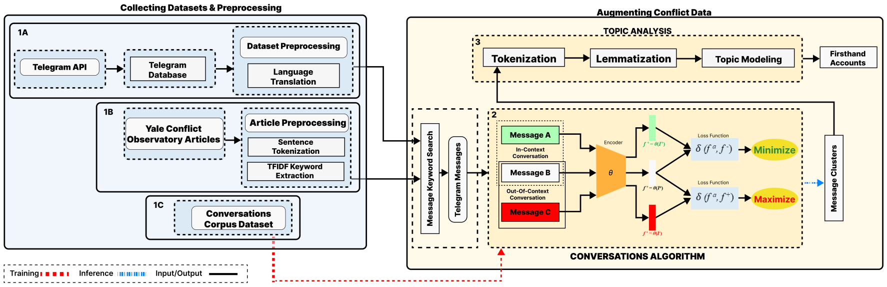

Fig. 1. Overview of our framework.

the Articles from the yale conflict observatory. In the Keyword Extraction stage, we utilize the Term Frequency-Inverse Document Frequency (TF-IDF) method to identify key terms in the articles.

3.4 Conversation Thread Analysis

The Conversation Thread Analysis phase uses language models' natural language processing capabilities to analyze conflict communication. It starts with Message Search, using previously identified key phrases to find relevant Telegram messages. This targeted approach ensures collected data relates directly to the studied conflict and comes from active event discussions. After refining extracted keywords by removing special characters and duplicates, a curated set of distinct keywords retrieves a subset of 60,523 relevant messages from the Telegram dataset for detailed analysis.

Algorithm 1 modified from [16], depicts our conversations extraction process for the Telegram dataset subset. We use a fine-tuned RoBERTa-base model [3, 8] for binary classification of sentence pairs as part of the same conversation thread. The model $f(s_{i1}, s_{i2}; \theta)$ is trained contrastively on sentence pairs $P = (s_{i1}, s_{i2})i = 1^n$ with labels $L = l_i i = 1^n$, where $l_i \in 0, 1$ indicates same-conversation membership. θ represents learnable model parameters. The final model achieved 74% accuracy on the classification task.

Our study optimizes parameters in a contrastive learning framework by minimizing Cross-Entropy loss, which evaluates binary classification accuracy by penalizing differences between predicted and actual labels. We use the AdamW optimizer, a variant of Adam, known for efficiently handling sparse gradients and adjusting learning rates based on gradient moments. Post-training, we generate token pairs from messages using a BertTokenizer. Our pre-trained model then predicts conversation membership for each pair by computing an output vector, transformed into probabilities via a softmax function. This approach ensures precise predictions and robust handling of diverse data inputs.

Where K is the number of possible classes (2: in the same conversation or not), we consider two sentences to be in the same conversation if the probability p_{ij} exceeds a chosen threshold θ. If p_{ij} indicates the messages are not in context,

Algorithm 1 Message Conversation Grouping

```
Require: pre-trained model $\mathcal{M}$, input data set $D = \{d_1, d_2, ..., d_n\}$, minimum matches $\delta$
Load pre-trained model $\mathcal{M}$
for each data $d$ in $D$ do
    Load data into generic datastore $DS$
    Initialize empty datastore $DS_g$ for grouped messages
    Initialize group id $g_{\text{id}} = 0$
    for each pair of messages $(m_i, m_j)$ in $DS$ do
        Extract embeddings for $m_i$ and $m_j$
        Generate prediction $p_{ij}$ using $\mathcal{M}$ on embeddings of $m_i$ and $m_j$
        if $p_{ij}$ signifies messages are not in context then
            Increment $g_{\text{id}} = g_{\text{id}} + 1$
        end if
        Add messages to respective group in $DS_g$
    end for
    Format 'message_count' in $DS_g$
    Filter groups in $DS_g$ based on $\delta$
    for each group $g$ in $DS_g$ do
        Save $g$ to an appropriate output
    end for
end for
```

a new cluster is created by incrementing the cluster id $c_i d = c_i d + 1$. Messages are then added to their respective clusters in DF_c, our clustered messages datastore. This method dissects the vast Telegram message corpus into distinct conversations using refined keywords, akin to targeted crowdsourcing. Given the data volume and natural language complexity, this automated approach is more practical and effective than manual analysis.

We use a trained conversations detection model to filter and group messages into topically coherent conversation threads, iteratively refining our data organization. Our algorithm systematically examines each message's relationship with all others to determine conversation membership. Identified conversations are grouped, and the process repeats for each ungrouped message. After all messages are evaluated and grouped, we compare the hash of each group sorted by messages_id to identify and remove any duplicate groups. This LLM-driven approach ensures efficient and accurate conversation clustering, crucial for handling the complexity and volume of our data.

Our model accommodates non-transitive relationships between messages, reflecting natural conversational structures. Ambiguities in message clustering can lead to inconsistencies, as illustrated in this scenario with three messages (m1, m2, m3): (1) m1 and m2 are in the same conversation, (2) m2 and m3 are not, but (3) m1 and m3 are. To address such challenges, our algorithm allows message replication across groupings, ensuring no conflicts arise and enabling messages to exist in multiple groupings as context demands.

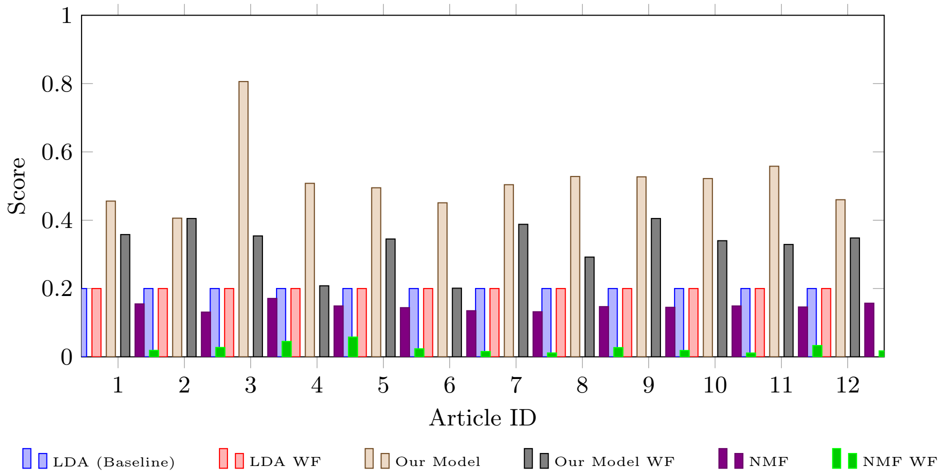

Fig. 2. Comparison of topic model scores for each article.

3.5 Clustering Conversations: Context vs. Topic

Our ’contextual topic clustering’ approach combines context-based and topic-based clustering methods to overcome their individual limitations. We leverage the language model’s classification for context clustering and reconstruct original conversation threads using message IDs and reply-to IDs for topic clustering. This method yields sub-corpora of related message groups connected by both context and topic. The resulting clusters contain messages from individuals discussing the same subject within the same conversation thread, ensuring both contextual relevance and semantic coherence. This approach provides a more comprehensive representation of conversations, capturing overall context and specific topics. It’s valuable for applications like sentiment analysis, opinion mining, and trend detection, offering a granular understanding of the discourse within the dataset.

3.6 Augmenting Conflict Data

Our final process involves topic analysis on conversation clusters using Latent Dirichlet Allocation (LDA), a generative probabilistic model. For a corpus of D documents, each document d with words w, and T topics with word distribution β_t, LDA finds: (1) topic distribution θ_d and (2) word topic assignment $z_{d,w}$. The generative process involves choosing $\theta_d \sim \text{Dirichlet}(\alpha)$ for each document, then for each word: $z_{d,w} \sim \text{Multinomial}(\theta_d)$ and $w_{d,w} \sim \text{Multinomial}(\beta_{z_{d,w}})$. Parameters α and β control topic and word mixtures. We apply a Coherence Model [11] to assess topic coherence, yielding a quantitative measure of semantic coherence. This process provides a structured understanding of conflict data, unveiling underlying themes and sentiments, forming the basis for our conclusions and recommendations.

3.7 Implementation details

We implement our LDA model using Python libraries NLTK, Gensim, and Pandas. The pipeline includes: text preprocessing (tokenization with NLTK’s regu-

Table 2. Evaluation of Framework & Articles

Top-Article Keyword	**New Insights**	**Sample(s)**
impact, climate, gastronomy	Findings on the extent of potential damage to Ukrainian cultural heritage	Ukraine **bombed** the museum and **civilians were targeted** and injured as well
vdv, defence, force, oblast, airborne	Exploring the causes and implications of the deteriorating situation in Kyiv	Drone footage of a tank of the 35th Ukrainian marine brigade **shelling Russian positions in close combat**. **Destruction of a large amount of** enemy equipment and manpower by Marines of the 36th Brigade

lar expression tokenizer and lemmatization with WordNet), converting text to bag-of-words representation using Gensim's dictionary class, training the LDA model (Gensim implementation with adjustable topic number), generating topics, and calculating coherence scores using Gensim's CoherenceModel. This score evaluates topic quality by measuring semantic similarity of high-scoring words within each topic. The implementation uses Gensim's internal handling of learning parameters like learning rate and epoch number, employing online stochastic inference for optimization and determining epochs based on model convergence.

4 Evaluation

This section assesses our framework's efficacy and validity on the Telegram dataset through both quantitative and qualitative analyses. We evaluate our model's conversation clustering performance using intrinsic dataset attributes, reconstructing conversations from message ID and reply-to ID. This method allows us to directly compare the true conversation structure with our model's predicted groupings, providing an accurate assessment of the model's performance. Our approach combines quantitative metrics with qualitative evaluation to ensure a comprehensive understanding of the framework's effectiveness.

Quantitative Analysis In our initial evaluation, we compared three topic modeling techniques: (1) Latent Dirichlet Allocation (LDA), a standard topic modeling algorithm; (2) Our Custom Model, which uses NLTK for preprocessing and an LDA-based approach; and (3) Non-Negative Matrix Factorization (NMF), an alternative topic modeling method. This assessment focused on keyword choice for topic modeling and our model's performance relative to traditional techniques.

To assess the impact of our preprocessing and keyword extraction stages, we compared the performance of the three techniques with and without these initial framework phases, labeling results without these stages as "Without Framework" (WF). We evaluated performance based on each model's ability to derive meaningful topics from Telegram conversations, using topic coherence scores as the metric. Higher coherence scores indicate greater semantic similarity among top-ranking words in a topic, reflecting the model's proficiency in capturing meaningful topics. This comparison helped us understand the specific influence of our preprocessing and keyword extraction on topic analysis outcomes.

Figure 2 compares the three topic modeling techniques for each article, with and without our proposed framework. Our custom model, using NLTK for preprocessing and an LDA-based approach, achieved an average coherence score of 0.5219 across all articles, compared to 0.2 for standalone LDA and 0.1462 for NMF. Using the formula Improvement = $\frac{\text{Our Model Score}-\text{Baseline Score}}{\text{Baseline Score}} \times 100\%$, our model showed a 158.13% improvement over the standalone LDA baseline: $\frac{0.5219-0.2}{0.2} \times 100\% = 158.13\%$.

Our custom model's 158.13% improvement over standalone LDA underscores the effectiveness of our preprocessing techniques. The graph in Figure 2 shows that excluding our framework's initial stages leads to comparable or diminished performance across all models. While standalone LDA sometimes lacked clarity and NMF showed inconsistencies, our NLTK-enhanced model excelled in providing nuanced insights into the Telegram dataset's geopolitical discourse. To optimize keyword count, we tested configurations ranging from 3 to 10 keywords, evaluating their impact using average coherence scores. This approach balanced comprehensive theme capture with avoiding redundancy. The results of this keyword count evaluation are presented in Table 3.

Table 3. Different Keyword Counts

Keywords	Average Score
3	0.4235
5	0.5219
10	0.4523

Table 4. Top Keywords per Article

Keywords
news, preservation, artifacts, restoration, und
light, darkness, instability, energy, soldier

Table 3 shows that using 4 keywords decreased the average coherence score by 15% (from 0.5219 to 0.4519) compared to 5 keywords, indicating missed crucial themes. Increasing beyond 5 keywords didn't significantly improve scores, with 6 keywords showing a marginal 2-3% decrease, suggesting potential noise introduction. The 5-keyword configuration emerged as optimal, achieving the highest average coherence score of 0.5219. This balance captured dataset nuances while maintaining thematic coherence and relevance, maximizing the effectiveness and clarity of our thematic analysis.

Qualitative Evaluation Following our model's quantitative improvement, we conducted a qualitative evaluation to explore nuanced thematic interpretations within the dataset. We identified 'top topics' from conversation threads for each article, assigned by ID. This interpretative analysis, presented in Table 2, provides a multi-dimensional view of the dataset's subtler narratives and dominant discourses. Top keywords served as indicators of main narratives: terms like 'russian', 'ukraine', and 'war' highlighted geopolitical conflict; 'health', 'covid', and 'vaccine' indicated public health discussions; while 'looting', 'heritage', and 'damage' revealed concerns about cultural preservation amidst global unrest.

This approach unraveled the complex thematic landscape within the conversations.

We extracted the top 5 keywords from topic modeling across each article's clusters (Table 4) for a more granular view of dominant narratives. Keywords like 'child', 'health', and 'covid' in Articles 5, 8, and 12 highlight the crisis's impact on children's well-being and public health. Terms such as 'looted', 'destroyed', and 'damage' in Articles 3, 4, and 10 underscore discussions about widespread destruction and cultural heritage issues. This combined quantitative and qualitative evaluation validates our framework's effectiveness and demonstrates its potential for extracting and analyzing meaningful topics from complex conversation datasets like Telegram, accurately identifying and interpreting conversation threads and their underlying themes.

5 Discussion

5.1 Beyond Ukraine & Future Works

The methodology and framework developed in this research, while tailored to the context of Ukraine, hold the potential for broad applicability in other regions and scenarios. The modular nature of our approach, which integrates data collection, message clustering, and contextual topic analysis, can be adapted to different datasets, languages, and cultural contexts. By refining search terms, adjusting the model parameters, or incorporating region-specific nuances, the approach can be generalized to study other conflict zones or areas of interest. Furthermore, the ethical considerations and biases identified in our research context provide valuable insights that can guide adaptations in other scenarios, ensuring rigorous and responsible research practices.

5.2 Limitations and Ethical Considerations

This research faces several limitations and ethical considerations. The use of specific search terms for data collection risks excluding relevant content, necessitating iterative refinement of queries and continuous data validation. Relying on Yale Humanitarian Research Lab's satellite articles, while reputable, may introduce inherent biases or focus areas that could influence result interpretation. Additionally, analyzing messages outside their original context risks misinterpretation or loss of nuanced meanings, highlighting the challenge of accurately understanding and presenting data in conflict communication research. Another limitation; Social media data may include misinformation or hate speech, partially mitigated by crowdsourcing also Keyword extraction sometimes yields irrelevant content, which could be addressed through human evaluation.

6 Conclusion

In this work, we introduced our framework for the crowdsourcing of secure social media conversations in conflict zones. We collected a dataset of Telegram posts to demonstrate the capability of our framework and draw new insights into the Russia-Ukraine conflict. Our framework outperforms baselines by 158.13% and captures 477 new conversational groups pertaining to key new insights on specific events in the ongoing conflict, such as the bombing of Ukrainian cultural heritage sites and health facilities. Our research showcases the utility of crowdsourcing firsthand account in conflict zones using secure social media conversations.

Acknowledgments. This research project and the preparation of this publication were funded in part by the NSF Grant No. 2245983.

References

1. College, N.W.: Types of intelligence collection - intelligence studies. https://usnwc.libguides.com/intelligence/studies (2023), accessed: 2023-07-02
2. Cornell University: Cornell ConvoKit: A Collection of Conversations from Wikipedia Talk Pages. https://convokit.cornell.edu/documentation/awry.html (2023), accessed: July 11, 2023
3. FacebookAI: roberta-base. https://huggingface.co/FacebookAI/roberta-base (2019)
4. Ge, Q., Hao, M., Ding, F., et al.: Modelling armed conflict risk under climate change with machine learning and time-series data. Nat Commun **13**, 2839 (2022). https://doi.org/https://doi.org/10.1038/s41467-022-30356-x
5. Gleditsch, K.S., Metternich, N.W., Ruggeri, A.: Data and progress in peace and conflict research. Journal of Peace Research **51**(2), 301–314 (2014), http://www.jstor.org/stable/24557423
6. Institute, F.P.R.: Understanding russia's invasion of ukraine. https://www.fpri.org/ (2022), accessed: 2023-07-02
7. Lab, Y.H.R.: Conflict observatory (2023), https://medicine.yale.edu/lab/khoshnood/, accessed: 2023-07-10
8. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
9. Lyzem: Lyzem - Privacy Friendly Search Engine. https://lyzem.com/ (2023), accessed: July 11, 2023
10. Ramos, J.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning. pp. 29–48. Citeseer (2003)
11. Rehurek, R.: Gensim: Topic modelling for humans. https://radimrehurek.com/gensim/models/coherencemodel.html (2022), accessed: 2023-06-13
12. Research, C.P.: Telegram becomes a digital forefront in the conflict - news feeds from fighting zones. https://blog.checkpoint.com/ (2023), accessed: 2023-07-02
13. Salehyan, I.: Best practices in the collection of conflict data. Journal of Peace Research **52**(1), 105–109 (2015), http://www.jstor.org/stable/24557521
14. Tsovaltzi, D., Judele, R., Puhl, T., Weinberger, A.: Leveraging social networking sites for knowledge co-construction: Positive effects of argumentation structure, but premature knowledge consolidation after individual preparation. Learning and Instruction **52**, 161–179 (2017). https://doi.org/https://doi.org/10.1016/j.learninstruc.2017.06.004
15. Zeitzoff, T.: How social media is changing conflict. The Journal of Conflict Resolution **61**(9), 1970–1991 (2017), http://www.jstor.org/stable/26363973
16. Zhang, Y., Wang, Z., Shang, J.: Clusterllm: Large language models as a guide for text clustering. ArXiv **abs/2305.14871** (2023). https://doi.org/10.48550/arXiv.2305.14871