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

Abanisenioluwa Orojo[§*], Pranish Bhagat[*], John Wilburn[†], Michael Donahoo[§*], Nishant Vishwamitra[§†]
[*]Baylor University, [†]University of Texas at San Antonio
{abanisenioluwa_oroj1, pranish_bhagat1, Jeff_Donahoo}@baylor.edu
nishant.vishwamitra@utsa.edu, john.wilburn@my.utsa.edu

***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.**

I. Introduction

The intricate fabric of today's global societal landscape is increasingly dominated by incidents of political instability, rising tensions, and outbreaks of violent conflicts that dramatically affect societies on multiple levels. In this age of rapidly escalating conflicts, exemplified by the situations such as the 2022 Russian invasion of Ukraine [1], gathering accurate firsthand account information from these volatile environments has never been more critical. As these complexities proliferate, the necessity for nuanced, comprehensive, and timely insight into on-the-ground situation in these conflict zones grows exponentially.

Traditionally, firsthand account intelligence gathering has relied primarily on conventional mechanisms such as satellite imagery, on-the-ground journalism, and information from diplomatic sources [2]. These methods, while effective, offer a limited lens through which to view the unfolding realities on the ground, since turbulent and dangerous situations in conflict zones prohibit surveying the public on the ground. As a result, they often provide a delayed, fragmented, or incomplete picture, lacking the ability to capture the full spectrum of human experiences and perspectives within the conflict zones.

The advent of the digital age, however, has ushered in a wealth of new data sources that have the potential to enrich our understanding of these conflict-ridden regions [3], [4]. In particular, secure social media such as Telegram [5] has emerged as a powerful tool [5], offering people on the ground to safely communicate first-hand accounts of the events on the ground, resulting in a constant stream of real-time, user-generated content that reflects a diverse range of viewpoints. Recent studies have shown that these secure social media platforms are being increasingly used by people in conflict zones to communicate vital firsthand account information [6]. Yet, with the availability of such vast and multifaceted data comes a considerable challenge: the effective extraction of meaningful insights from these complex data streams [7], [8].

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 [9]. While relevant firsthand account reports by users are present, they are often lost among other conversations, such as criticizing the entities involved, or general expressions of emotions about the crisis situation. Techniques that can capture these multiple interweaving conversations are needed, which would enable extracting meaningful interactions and tease out the key themes [10], [11].

In this work, we propose a new framework to crowdsource firsthand account information from users of secure social media in conflict zones. Since users of secure social media platforms, such as Telegram share vital firsthand account information that cannot be sourced through traditional means, we aim to leverage these interactions to augment open-source information. Our framework first uses TF-IDF [12] to extract key phrases that pertain to specific topics regarding a conflict. Since messages pertaining to these extracted phases could be of several types, ranging from valid firsthand account information to expressions of emotions to condemnations of the involved entities, our framework addresses this key challenge of organizing the messages pertaining to the extracted phrases based on a novel conversations extraction algorithm based on RoBERTa [13], an LM. By organizing the messages into conversations, our framework can identify those conversations that are reporting firsthand account, and separate them from other conversations that are not relevant to the task of crowdsourcing, such as expressions of emotions.

We then use our framework to answer 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?

[§]Corresponding authors

To this end, we focus on the Russia-Ukraine conflict[1] and first collect a dataset of 250,000 Telegram (*i.e.*, a social media platform popularly used for secure communication) messages[2] from the Russia-Ukraine conflict. Our framework addresses RQ1 by leveraging our novel conversations algorithm to identify the conversations that are pertaining to firsthand account information in conflict zones. Our conversations-based framework outperforms baselines by 158.13%, demonstrating the effectiveness of our approach, and the crucial importance of conversations in this problem. To address RQ2, we run our framework on our dataset of Telegram posts, and capture 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.

II. Related Works

Our exploration of the scholarly landscape encompasses four primary domains within conflict research: the application of social media in conflict, methodologies for conflict data collection, the use of empirical data in conflict research, and the detection of human rights violations through social media.

A. Social Media's Role in Conflict Dynamics

Social media's influence in modern conflicts is significant, serving as a critical communication tool for actors like politicians, insurgents, and protestors. Its impact in conflict situations encompasses four key aspects: lowering communication costs, accelerating information dissemination, prompting strategic adaptation in response to technology changes, and offering new data that shapes conflicts [10]. Social media also fuels conflict through disinformation campaigns, electoral manipulation, and online extremist recruitment, thus exacerbating conflict drivers and polarization [14]. Recognizing these challenges, peacebuilding efforts are increasingly incorporating a digital perspective. Notably, Mercy Corps developed a framework based on case studies in Ethiopia, Iraq, Myanmar, and Nigeria to counteract social media's weaponization and promote cohesion [14]. This underscores the importance of comprehensively understanding social media's complex role in both magnifying and resolving conflicts, relevant for scholars, policymakers, and peacebuilding practitioners.

B. Conflict Data Collection

In the field of conflict studies, data collection is a paramount challenge. Salehyan (2015) highlights the importance of data collection strategies that prioritize accuracy, consistency, and replicability [11]. An additional best practice that is highlighted is the inclusion of "non-events" or periods of peace, in conflict data. Salehyan also stresses the importance of intercoder reliability, which is especially relevant in the context of social media data collection, where the interpretation of the text and other media can be highly subjective [15].

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

[2]Our dataset will be made publicly available.

C. Empirical Data in Conflict Research

The scholarly works of Gleditsch (2020) and others have played an instrumental role in driving this shift, highlighting the fundamental importance of high-quality, disaggregated data for advancing our understanding of conflict processes [16] [17]. Such data sets allow researchers to delve into the specifics of conflict, examining factors such as the types of conflict, the actors involved, and the strategies they employ. Moreover, the advent of 'big data' has ushered in a new era of conflict research, pushing researchers to rethink how data should be aggregated for querying and evaluating specific theories. The challenges in collecting high-quality data, particularly in conflict research, are significant, given the difficulty in observing every conflict-related event and the reliance on varied sources for information [18]. Recent discussions in the conflict studies community have further emphasized the need for accurate, replicable, and interoperable data collection methods [19].

D. Social Media-Based Human Rights Violations Detection

In recent years, the use of social media platforms for detecting human rights violations in conflict zones has gained attention. Nemkova et al. (2023) [20], [21] have highlighted the effectiveness of social media in uncovering evidence of atrocities, such as the use of chemical weapons in Syria or extrajudicial killings in Myanmar. However, despite the potential benefits, there are several challenges associated with using social media for detecting human rights violations. A key challenge is the disorganized nature of information on social media, which can be overwhelming, making it difficult to identify and prioritize relevant content for analysis. Furthermore, the reliability and verification of social media data, the need for admissibility guidelines for legal contexts, and the risk of underreporting due to disparities in access to technology are significant concerns in this field [22].

III. Methodology

A. Data Collection

To achieve our objectives, we first build a rigorous data collection process to collect a dataset of secure social media messages from Telegram. We also collect a dataset of open-source articles that have been published via traditional reporting methods to demonstrate how our framework can augments such data with information from the source. Additionally, we use an existing dataset of social media conversations to train our LM algorithm [23]. Thus we collect data into two main streams: satellite articles and Telegram messages.

1) **Satellite Articles:** Our first stream of data is derived from satellite articles sourced from a traditional journalistic source called the Conflict Observatory by the Yale Humanitarian Research Lab (HRL) [24]. This organization is well-known for using traditional journalistic methods such as satellite imagery for comprehensive and unbiased reporting on conflict zones across the globe. The articles obtained from these sources are used as sources of traditional information that need to be augmented using crowdsourcing from conflict zones. The thirteen articles

TABLE I
SUMMARY OF ARTICLES AND DISCUSSED TOPICS

Article ID	Article Name	Topics Discussed	Date
1	Ukrainian Cultural Heritage Potential Impact Summary	impacts on Ukrainian cultural heritage, climate and gastronomy.	May 2022
2	Extrajudicial Detentions and Enforced Disappearances in Kherson Oblast	Illegal detentions and disappearances in Kherson Oblast.	Nov 2022
3	Ukraine's Crop Storage Infrastructure: Post-Invasion Damage Assessment	Damage inflicted on Ukraine's crop storage infrastructure	Sep 2022
4	Damage Assessment of Health and Educational Facilities in Sievierodonetsk Raion, Ukraine	The damage to health and educational facilities in Sievierodonetsk Raion.	Jul 2022
5	Evidence of Widespread and Systematic Bombardment of Ukrainian Healthcare Facilities	The systematic bombardment of Ukrainian healthcare facilities	May 2022
6	Russia's Systematic Program For The Re-education And Adoption of Ukraine's Children	Russia's program for re-educating and adopting Ukrainian children.	Feb 2023
7	Mass graves at Pishchanske Cemetery in Izyum	Discovery of mass graves in Izyum.	Mar 2023
8	Population Displacement and Return in Ukraine	Population displacement and return trends in Ukraine.	Sep 2023
9	Rapid Report: Kherson Regional Art Museum Reported Looting Event	Looting of the Kherson Regional Art Museum.	Nov 2022
10	Potential Damage to Ukrainian Sites	Damage to Ukrainian cultural sites.	Feb 2023
11	Analysis of Damage to Ukrainian Cultural Heritage Sites	Damage to Ukrainian cultural heritage sites.	Jun 2022
12	Mapping Russia's Detention Operations in Donetsk Oblast	Detention operations by Russia in Donetsk Oblast.	Aug 2022
13	Kyiv Falling into Darkness	Instability and decreased light production in Kyiv.	Nov 2023

from the Yale Conflict Observatory see Table I, offer a comprehensive overview of the Russia-Ukraine conflict from May 2022 to 2023. The topics covered by these articles span a variety of critical issues related to the conflict .

A significant portion of the articles focuses on impact assessments in different areas. These include assessments of potential and actual damages to Ukrainian cultural heritage sites, crop storage infrastructure, and health and educational facilities. The reports include detailed evaluations of infrastructural damage, the effects on local communities, and the broader implications for Ukraine's cultural and historical legacy. Other articles delve into more specific issues, such as the systematic bombardment of Ukrainian healthcare facilities, highlighting potential violations of international law. These pieces document and analyse patterns of attacks on critical infrastructure, exploring their widespread and potentially systematic nature. Additionally, the reports cover human rights concerns, such as extrajudicial detentions and enforced disappearances in specific regions of Ukraine. They provide an in-depth look at these illegal practices, evidencing grave human rights violations within the conflict. Furthermore, the articles explore the social consequences of the conflict, including the displacement and subsequent return of populations within Ukraine.

2) **Telegram Messages:** The second stream of our data is extracted from Telegram, a widely-used secure messaging platform with a broad user base and an active presence in the regions of interest. Telegram's substantial usage in conflict zones makes it an invaluable resource for understanding the ground reality and collecting first-hand accounts of the situation. Using the service provided by Lyzem (A Telegram Search Engine) [25], we focused our data collection efforts on group chats centered in the Ukraine region. To refine our data collection, we employed key search terms, such as "Ukraine", "Russia", "war", and other related terms. We used Google Trends to expand our keyword dataset, taking the top 25 Related topics and Related queries each for the foundational words like "Ukraine", "Russia", "war", "conflict", and "invasion". This scoping helped to ensure that the data collected was specifically related to the conflict under investigation. TableII depicts the result of our data collection process. Overall, we were able to collect a total of 258,101 messages from the start of our collection in 2020 up to 2023. A significant surge in the number of messages can be observed in 2022 when the conflict intensified, accumulating to 150,083 messages, and this continued into 2023 with a further 107,351 messages recorded. The messages downloaded from these group chats form a rich corpus of data, offering insights into the perceptions, sentiments, and discussions surrounding the conflict.
3) **Social Media Conversations Dataset:** Our work leveraged the Social Media Conversations Dataset available from Cornell University's ConvoKit. This comprehensive collection encapsulates conversations from various online platforms, with an emphasis on identifying early signs of conversational failure. From this rich dataset, we formulated our training dataset, which comprises pairs of messages and a corresponding label denoting whether the pair belongs to the same conversation or not. This refined dataset by grouping message pairs that fell within the same conversation, in contrast to pairs that did not share the same conversational context.

TABLE II
TELEGRAM MESSAGES COLLECTED PER YEAR.

Year	Count of Messages
2020	116
2021	551
2022	150,083
2023	107,351

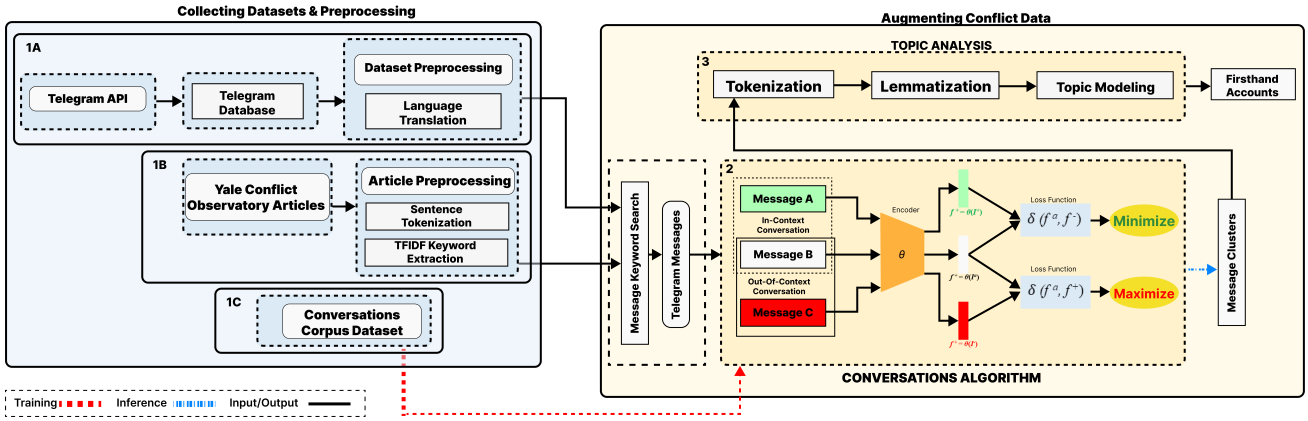

Fig. 1. Overview of our framework.

B. Overview of Our Approach

Our framework broadly consists of five main components.

1) Data Acquisition: The framework ingests two types of data - conflict-related articles from the Yale Conflict Observatory that require firsthand account analysis, and a large dataset of Telegram messages pertaining to the conflict.
2) Preprocessing & Keyword Extraction: The acquired article data undergoes preprocessing to clean and prepare it for analysis. TF-IDF is then applied to extract the most relevant keywords that represent the primary topics and focus areas of each article.
3) Message Searching: The extracted keywords are used to identify and collate a cluster of relevant messages from the telegram dataset that correspond closely to the keywords.
4) Conversation Modeling & Thread Identification: The resultant clusters of telegram messages are then fed into an LM-based conversation model. This model's primary task is to identify conversation threads.
5) Topic Analysis: The final step involves performing topic analysis on the identified conversation threads to extract firsthand account. By categorizing and analyzing the themes or subjects that these threads revolve around, we can pinpoint the key topics of discussion within the conflict data.

C. 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 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. The TF-IDF is calculated using the following formulas:

1. Term Frequency (tf):

$$tf(t,d) = \frac{\text{Count of term } t \text{ in document } d}{\text{Total terms in document } d} \quad (1)$$

2. Inverse Document Frequency (idf):

$$idf(t,D) = \log\left(\frac{\text{Total documents in corpus } D}{\text{Documents containing term } t}\right) \quad (2)$$

3. TF-IDF Score:

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D) \quad (3)$$

D. Conversation Thread Analysis

The Conversation Thread Analysis phase employs a structured approach using the natural language processing capabilities of language models (LMs) to analyze conflict communication. This phase begins with the Message Search stage, where key phrases previously identified are used to search for relevant Telegram messages. This targeted search ensures the data collected is directly related to the conflict being studied, and originates from active discussions about the event. After refining the extracted keywords by removing special characters and duplicates, a curated set of distinct keywords is used to retrieve a pertinent subset of 60,523 messages from the Telegram dataset for in-depth analysis.

The steps in the conversations extraction algorithm are depicted in Algorithm 1 which is used on the Telegram dataset subset. We utilize a fine-tuned RoBERTa-base model [13], [26], a transformer-based LM that is pre-trained on a large corpus of text data. Our task is binary classification, determining whether two sentences are part of the same conversation thread. The model is trained in a contrastive manner using pairs of sentences with labels indicating whether they belong to the same thread. Let the set of all sentence pairs be denoted as $P = (s_{i1}, s_{i2})i = 1^n$ and the corresponding labels as $L = l_i i = 1^n$, where $l_i \in 0, 1$ signifies whether sentences s_{i1} and s_{i2} are in the same conversation (1) or not (0). The RoBERTa model, represented as a function $f(s_{i1}, s_{i2}; \theta)$, takes a sentence pair and outputs a prediction. θ represents the model parameters which are learned during the training process. The final fined model achieved an accuracy of 74% on the classification task.

In our study, we aim to optimize parameters in a contrastive learning framework by minimizing the Cross-Entropy loss. This loss function evaluates the accuracy of our predictions

Algorithm 1 Message Conversation Grouping

Require: pre-trained model $\mathcal{M}$, input data set $D = \{d_1, d_2, ..., d_n\}$, minimum matches δ

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 δ
  for each group g in DS_g **do**
    Save g to an appropriate output
  end for
end for

for binary classification, where it penalizes the difference between predicted and actual labels in terms of the probability assigned to the correct label. The optimization of these parameters is conducted using the Adam optimizer, a method known for its efficiency in handling sparse gradients on noisy problems. AdamW, a variant we employ, adjusts the learning rate for each parameter based on estimates of first and second moments of the gradients, enhancing the stability of our optimization process. After training the model, we start by generating token pairs from the messages using a BertTokenizer function. For each message pair, our pre-trained model predicts whether they belong to the same conversation by computing an output vector. This vector is then transformed into probabilities using a softmax function, indicating the likelihood of the messages being part of the same conversation thread. This methodical approach ensures that our model is both precise in its predictions and robust in handling diverse data inputs.

Where K is the number of possible classes (in this case, 2: in the same conversation or not). We say that two sentences are in the same conversation if the probability p_{ij} is greater than a chosen threshold θ. If p_{ij} signifies that the messages are not in context, a new cluster is created by incrementing the cluster id $c_i d = c_i d + 1$. The messages are then added to their respective clusters in DF_c, our datastore for clustered messages. This method allows us to dissect the vast corpus of Telegram messages using refined keywords into distinct conversations. This is akin to a targeted crowdsourcing data collection effort, where we seek out and collect relevant information from these group chats, thereby enhancing the depth of our data pool. Given the large volume of data and the complexity of natural language, manual analysis is neither practical nor effective. We use the trained conversations detection model to filter and group the messages into conversation threads. This LLM-driven grouping is done based on topical coherence, where each thread represents a discussion on a specific topic or sub-topic. The construction and refining of these conversation threads are done iteratively, continually improving our data organization.

Our algorithm employs a systematic approach to group in-conversation messages. The process begins by taking each individual message and examining its relationship with all other messages to determine if they are part of the same conversation. When a set of messages is identified as a conversation, they are grouped together. The model then moves on to the next ungrouped message and repeats the process. Once all messages have been evaluated and grouped, the we conducts a final step where we compares the hash of each group sorted by messages_id. This comparison is crucial for identifying and removing any duplicate groups that may have been formed during the clustering process.

Our model acknowledges and intentionally accommodates non-transitive relationships between messages, reflecting natural conversational structures. Ambiguities in message clustering can occasionally arise, leading to potential inconsistencies in grouping messages into conversations. To illustrate, consider the following scenario: Given three messages: m1, m2, and m3, the model makes the following predictions:

1) m1 and m2 are in the same conversation.
2) m2 and m3 are not in the same conversation.
3) m1 and m3 are in the same conversation.

Such situations can present challenges in accurately clustering messages. To address such situations, our algorithm allows for message replications across groupings. This approach ensures that there's no conflict when similar scenarios arise, and each message can exist in multiple groupings if the context so demands.

E. Clustering Conversations: Context vs. Topic

Clustering messages into conversations can be approached from various angles, with context and topic being the primary dimensions. Context-based clustering focuses on the immediate surroundings of a message, accounting for the sequence, timing, and inter-relationships between messages. On the other hand, topic-based clustering emphasizes the semantic content, grouping messages that discuss similar themes or subjects. However, solely relying on either method can be limiting. Context clustering might group unrelated messages that occur in sequence, while topic clustering can scatter related messages across different groups if they touch upon multiple themes. To overcome these limitations, we introduce the concept of 'contextual topic clustering.' This method combines the strengths of both approaches, ensuring that messages are grouped based on the themes they discuss while also considering the context in which they appear. By integrating both dimensions, our framework offers a more nuanced and accurate conversation clustering solution.

Our contextual topic clustering approach leverages the language model's classification to obtain the context clustering.

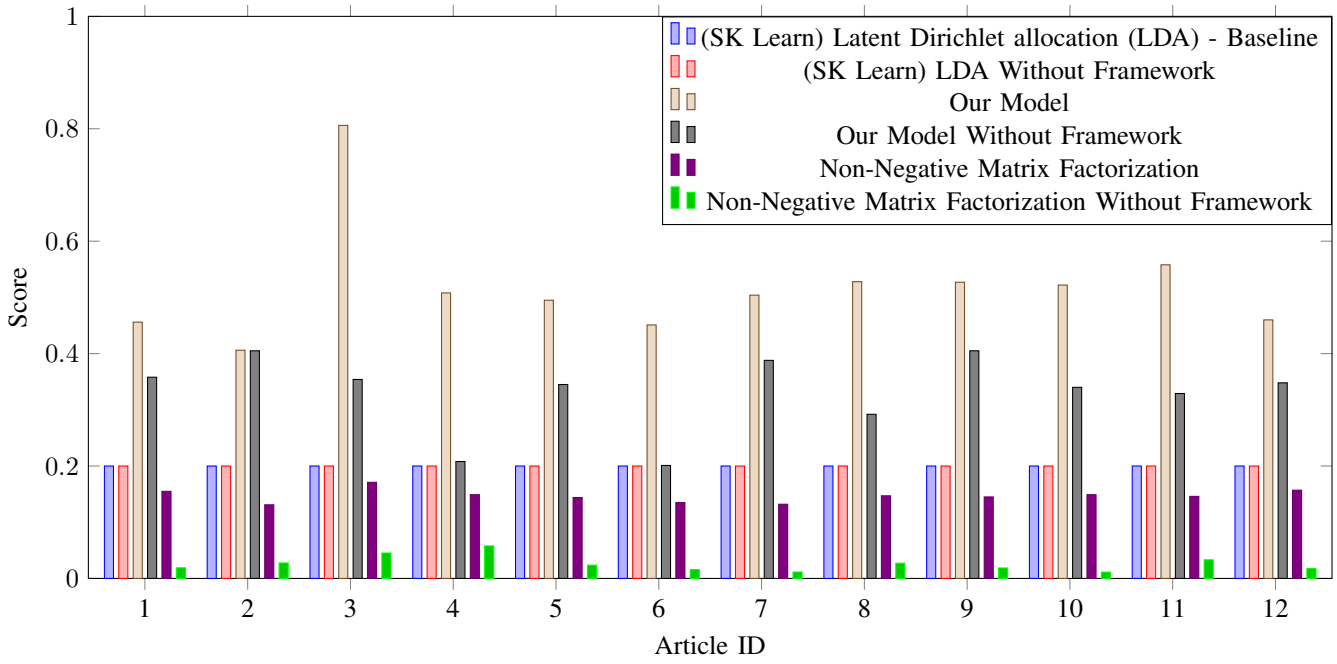

Fig. 2. Comparison of topic model scores for each article.

We then reconstruct the original conversation threads by utilizing the message IDs and reply-to IDs present in the dataset. This process yields a sub-corpus of related message groups that are not only connected by their context, i.e., being part of the same conversation, but also by their topic. The resulting clusters contain messages where individuals are engaging in discussions about the same subject, as evidenced by their replies to one another within the conversation thread. By combining the context derived from the language model and the inherent structure of the conversations based on the reply relationships, we create a clustering solution. This approach ensures that the resulting clusters are both contextually relevant and semantically coherent. The context clustering guarantees that messages within a cluster are part of the same conversational flow, while the topic clustering based on the reply structure ensures that the messages within a cluster are focused on the same theme or subject matter. This contextual topic clustering method provides a more comprehensive and accurate representation of the conversations taking place in the dataset. It captures not only the overall context of the discussions but also the specific topics being addressed within each conversation. This rich information can be invaluable for various applications, such as sentiment analysis, opinion mining, or trend detection, as it allows for a more granular understanding of the discourse taking place within the dataset.

F. Augmenting Conflict Data

The final process, Augmenting Conflict Data, involves performing a topic analysis on the conversation clusters identified by our conversations LM. We carry out our topic analysis using a Latent Dirichlet Allocation (LDA) model, a method for topic modeling in text data. The LDA model is a generative probabilistic model, and it works under the assumption that each document within the corpus is a mixture of a set number of topics, and each word in the document is attributable to one of the document's topics.

Given a corpus with D documents, and each document d is represented as a list of words w. If we suppose there are T topics in the corpus and each topic t is represented by a distribution over words β_t, LDA aims to find two things for each document d:

1) A distribution over topics θ_d
2) A topic assignment for each word $z_{d,w}$

These are computed using the following generative process, where the topics β_t and the per-document topic distributions θ_d are sampled from Dirichlet distributions:

For each document d:

1) Choose $\theta_d \sim \text{Dirichlet}(\alpha)$
2) For each word w in d:
   a) Choose a topic $z_{d,w} \sim \text{Multinomial}(\theta_d)$
   b) Choose a word $w_{d,w} \sim \text{Multinomial}(\beta_{z_{d,w}})$

The parameters α and β are corpus-level parameters that control the mixture of topics in a document and the mixture of words in a topic respectively.

Subsequently, we utilize the LDA model to discern the chief topics in the conversation threads. The number of topics is a tunable parameter, and each topic is represented by a set of words. To assess the coherence of the topics discerned by the LDA model, we apply a Coherence Model [27]. This model yields a quantitative measure of the topics' semantic coherence, which aids in evaluating the quality of our topic model. Following this, we identify prevalent and significant themes across these clusters. The outcome of this process is a structured and in-depth understanding of the conflict data. It not only mirrors the main discussion topics but also unveils underlying themes and sentiments in the conflict discourse.

This comprehensive understanding forms the backbone of our conclusions and recommendations, offering a nuanced and thorough analysis of the conflict situation.

G. Implementation details

We implement our Latent Dirichlet Allocation (LDA) model in Python using libraries such as NLTK, Gensim, and Pandas. The pipeline includes several key steps: preprocessing the text, converting the text into a bag-of-words representation, training the LDA model, generating the topics, calculating the coherence score for each topic, and then consolidating all the data.

For preprocessing the text, we first tokenize the text into individual words using a regular expression tokenizer from the NLTK library. This tokenizer is designed to capture alphanumeric words, effectively ignoring punctuation. After tokenization, we perform lemmatization using WordNet's lemmatizer to reduce words to their root form.

We then convert the processed text into a bag-of-words representation using Gensim's dictionary class. This representation is used as input to the LDA model. We use Gensim's implementation of the LDA model. The number of topics is a hyperparameter that is passed to the function and can be adjusted based on the desired granularity of topics. After training, the model can be used to generate the top words for each topic.

We calculate the coherence score for each topic using Gensim's CoherenceModel class. This score is a measure of the semantic similarity of the high-scoring words within each topic, providing a way to evaluate the quality of the learned topics. In terms of learning parameters, this implementation does not explicitly specify values for parameters such as learning rate or the number of epochs. These parameters are handled internally by the Gensim LDA model, which uses a variant of online stochastic inference for optimization and determines the number of epochs based on the convergence of the model.

IV. Evaluation

A. Evaluating our Framework on the Telegram Dataset

This section aims to assess the efficacy and validity of our proposed framework on the Telegram dataset from two distinct perspectives: a quantitative analysis and a qualitative evaluation.

B. Performance Metrics for Message Conversation Clustering

To evaluate the performance of our model's conversation clustering capability, we rely on the intrinsic dataset attributes. Specifically, using the message ID and reply-to ID, we can reconstruct the conversation from the dataset. Once reconstructed, we check if the corresponding messages appear in the same group as predicted by our model. This method provides a direct comparison between the true conversation structure and our model's predictions, ensuring an accurate assessment of the model's grouping performance.

1) Quantitative Analysis: In the first part of our evaluation, we assessed the choice of keyword for topic modeling and the performance of our model in topic analysis relative to other traditional techniques. Specifically, we conducted an analysis among three distinct topic modeling techniques:

1) **Latent Dirichlet Allocation (LDA)** - A well-established topic modeling algorithm.
2) **Our Custom Model** - Utilizes the Natural Language Toolkit (NLTK) for preprocessing and incorporates an LDA-based approach for topic modeling.
3) **Non-Negative Matrix Factorization (NMF)** - Another method for topic modeling.

To understand the impact of our preprocessing and keyword extraction stages, we compared the performance of these techniques with and without these initial phases of our framework. The results without these stages are termed as "*Without Framework*". This comparison helps us understand the distinct influence of our preprocessing and keyword extraction on the topic analysis outcomes.

Performance was gauged based on each model's capability to derive meaningful topics from the Telegram conversations. We employed *topic coherence scores* as the metric, which measures the quality and clarity of the derived topics. A higher coherence score suggests a greater semantic similarity among the top-ranking words in a topic, indicating the model's proficiency in capturing meaningful topics.

Figure 2 depicts a plot that contrasts the three topic modeling techniques for each article, both with and without our proposed framework. Our custom model, which leverages NLTK for preprocessing and an LDA-based approach for topic modeling, achieved an average coherence score of 0.5219 across all articles. In comparison, the standalone LDA and NMF models achieved average coherence scores of 0.2 and 0.1462, respectively. To calculate the percentage improvement, we used the following formula:

$$\text{Improvement} = \frac{\text{Our Model Score} - \text{Baseline Score}}{\text{Baseline Score}} \times 100 \tag{4}$$

Using the standalone LDA model as the baseline, our model achieved a 158.13% improvement:

$$\text{Improvement} = \frac{0.5219 - 0.2}{0.2} \times 100 \tag{5}$$

This substantial improvement highlights the merits of our comprehensive preprocessing techniques and the resultant high coherence scores. The graph further accentuates that excluding the initial stages of our framework leads to comparable or diminished performance across all models. While our custom model excelled, standalone LDA and NMF exhibited certain limitations. LDA sometimes yielded topics that lacked clarity, whereas NMF demonstrated inconsistencies in its results. These observations validate the effectiveness of our NLTK-enhanced model in delivering nuanced insights into the geopolitical discourse present within the Telegram dataset, especially when integrated into our holistic framework.

In determining the optimal number of keywords for our analysis, we conducted extensive testing with different key-

TABLE III
EVALUATION OF FRAMEWORK & ARTICLES

Article ID	Top-Article Keyword	New Insights	Sample(s)
1	**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
2	russian, report, supermarket, open, ukraine	Evidence of ongoing human rights abuses in the region	• Most of the severely weakened VDV units were dedicated to the defence of the Russian-held **territory** west of the Dnipro River in Kherson Oblast (province)
3	world, million, russia, ukraine	Insight into the implications for food security in Ukraine and potential international impact	• Artillery hits a **building** with 5 Russians inside • Russian **artillery lands on a library** in Mariupol
4	dive, russian, att	Evidence of indiscriminate and persistent bombardment by Russia-aligned forces	• By attacking Ukraine's **critical civilian infrastructure**, the Russian army clearly intends to undermine industrial production, disrupt transport
5	child, health, defense, **wounded**	Highlights of the catastrophic implications on healthcare in the region	• Wagner militants are wearing Ukrainian uniforms making those **rumors** a reality • Nine streets came under fire. As a result of the shelling, three civilians were wounded: a 16 years old boy and two men. Presently, all the **injured** are in hospital.
6	health, defense, **culture**, russian	Potential threats to Ukrainian cultural heritage	• They want to **"enrussify"** and for that Museums and bookshops need to be closed or repurposed
7	ukrainian, **children**, education	The strategies and implications of Russia's re-education and adoption programs	• They have been **kidnapping donezk children** • I gotta rest but for fuck sake **quit abusing children**. Please, please, please quit abusing children
8	earth, grave, post, Izyum	Uncovering evidence of mass graves and potential human rights violations	• There is no sewage, water, electricity, or gas in the city. Ninety percent of houses in the city are damaged or destroyed. **Corpses on the streets have begun to decompose because of the heat**, which can provoke epidemics of various diseases. People stand in huge queues for humanitarian aid, which is barely enough.
9	ukraine, regions, wondering, **refugee**	Understanding the implications of population displacement and the potential for return	• Mercenaries from the Middle East were in Volchansk, **Refugees** from the Kharkov region told us.
10	building, **looted**, russian, destroyed	Examination of looting patterns and implications on cultural heritage	• Russian soldiers are **looting** Ukrainian smart museums and exhibits for valuable items
11	russian, dive, **Donetsk, oblast**	Insights into Russia's detention strategies and human rights violations	• Volodymyrivka, Donetsk Oblast, about 7km East of Soledar - Ukrainian artillery fire against Russian positions is recorded by drones flying overhead.
12	**child**, health, defense, eat, covid	Unveiling the crisis's direct and indirect impact on children's health	• **Donezk children made Ukraine cold**. The children are not ugly. They are mutts • More than 15,000 people are constantly staying in the Kyiv metro Among them **there are 84 infants, 413 children of preschool and school age.**
13	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

word counts and evaluated their impact using average coherence scores. This approach aimed to find a balance between capturing a comprehensive range of themes and avoiding redundancy or over-specificity in the keyword set. We tested keyword counts ranging from 3 to 10 and calculated the average coherence score for each configuration. The results of this evaluation are presented in

Table IV.

As evident from Table IV, using 4 keywords resulted in a noticeable decrease in the average coherence score compared to the 5-keyword setup. Specifically, the performance dropped by approximately 15%, from 0.5219 to 0.4519. This suggested that crucial thematic elements were being missed, limiting the depth and richness of our analysis. Conversely, increasing the keyword count beyond 5 did not yield significant

TABLE IV
EVALUATION OF DIFFERENT KEYWORD COUNTS

Number of Keywords	Average Coherence Score
3	0.4235
4	0.4519
5	0.5219
6	0.5087
7	0.4952
8	0.4783
9	0.4641
10	0.4523

improvements. In fact, there was a marginal decrease in the average coherence score when using 6 keywords, around 2-3%, compared to the 5-keyword setup. This hinted at the introduction of noise or irrelevant information, as the additional keywords tended to make the representation overly specific and less focused. After this thorough evaluation, 5 keywords emerged as the optimal choice. This number provided a balance, ensuring that the keywords were sufficiently diverse to capture the dataset's nuances while maintaining a high level of thematic coherence and relevance. The 5-keyword configuration achieved the highest average coherence score of 0.5219, indicating its effectiveness in capturing meaningful and interpretable topics. Hence, the decision to use 5 keywords was driven by a desire to maximize the effectiveness and clarity of our thematic analysis.

2) Qualitative Evaluation: After establishing our model's quantitative improvement, we then conducted a qualitative evaluation to investigate the nuanced thematic interpretations within the dataset. We identified the 'top topics' from the conversation threads produced by our model. Each article was identified using an assigned ID and evaluated based on the 'top topics' representative of its main themes. The interpretative nature of this evaluation provides a multi-dimensional snapshot of the dataset, unraveling subtler narratives and dominant discourses in the conversations.

The table (Table III) presents the results of the qualitative analysis. We identified top keywords for each unique article in the corpus, which served as indicators of the main narratives in the dataset. For instance, terms like 'russian', 'ukraine', and 'war' highlighted a dominant discourse around geopolitical conflict. Recurring themes around public health issues were evident with the frequent appearance of 'health', 'covid', and 'vaccine'. The presence of specific keywords like 'looting', 'heritage', 'damage' revealed narratives around cultural preservation, indicating discussions and concerns around the protection of cultural assets amidst global unrest.

To further analyze the content of the conversation threads, we extracted the top 5 keywords from the topic modeling across each articles clusters, as shown in Table V. These keywords offer a more granular view of the dominant narratives and recurring themes within the firsthand accounts. For instance, the frequent appearance of terms like child', health', and covid' in Articles 5, 8, and 12 highlights the significant impact of the crisis on children's well-being and public health. The presence of keywords such as looted', destroyed', and damage' in Articles 3, 4, and 10 underscores the widespread destruction and cultural heritage issues discussed in the conversations.

This combination of quantitative and qualitative evaluation not only validated the effectiveness of our framework but also underscored its potential application in extracting and analyzing meaningful topics and narratives from complex conversation datasets like Telegram. The findings corroborate the utility of our proposed framework in accurately identifying and interpreting conversation threads and their underlying themes.

Article ID	Keyword 1	Keyword 2	Keyword 3	Keyword 4	Keyword 5
1	news	preservation	artifacts	restoration	und
2	super market	debatable	human-rights	unlawful	turkish
3	vdv	post-invasion	damage	always	brigade
4	impact	sievier-odonetsk	medical	want	defense
5	europe	mariupol	war-crimes	systematic	ukraine
6	systematic	redd	adoption	kyiv	indoct-rination
7	building	permit	show	destroyed	casualties
8	covid	health	refugee	redd	migration
9	earth	arthur	post	flat	start
10	ukraine	impact	conservation	russian	destruction
11	world	million	sites	operations	assessment
12	child	health	covid	vaccine	org
13	light	darkness	instability	energy	soldier

TABLE V
TOP 5 KEYWORDS FROM EACH ARTICLE

V. DISCUSSION

A. Generalizability Beyond Ukraine

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.

B. Contextualizing Our Approach within the Landscape of Social Media-Based Conflict Research

The exploration of methodologies leveraging social media data for insights into conflict zones presents a spectrum of strategies, each with unique advantages and challenges. In this subsection, we contextualize our approach within the broader landscape of social media-based conflict research by comparing our methodology with that employed in "Detecting Human Rights Violations on Social Media during the Russia-Ukraine War" [28].

Data Collection and Processing Mechanisms: Our research utilizes a robust dataset of over 250,000 Telegram messages pertinent to the Russia-Ukraine conflict. We introduce a deep learning algorithm founded on a language model (LM) to unveil conversational contextual groupings within social media messages, thereby enhancing conventional firsthand account-gathering practices with crowdsourced information. In contrast, Nemkova et al. [28] employ ChatGPT for data retrieval, navigating through the model's variability and ethical considerations in its response generation. Their methodology included the generation of 510 positive examples, which,

after a meticulous review to exclude irrelevant instances, were incorporated into their training data.

Analytical Approaches: Our research is fundamentally rooted in utilizing an LM to discern contextual groupings within the data, although the extracted snippet does not provide comprehensive details on the analytical nuances and subsequent steps following the primary data processing. Nemkova et al. [28], on the other hand, navigate through the challenges presented by ChatGPT, addressing its instability and the ethical considerations emerging during data retrieval. Their methodology suggests a structured approach wherein the generated responses undergo a review process to ensure relevance and accuracy, albeit further details would be requisite for a comprehensive understanding.

By contextualizing our approach within the broader landscape of social media-based conflict research, we highlight the unique aspects of our methodology while acknowledging the shared goal of leveraging social media data to gain insights into conflict zones. This comparative analysis underscores the diversity of strategies employed in this domain and the ongoing efforts to navigate the challenges and opportunities presented by these novel data sources.

C. Limitations and Ethical Considerations

In this research, several limitations and ethical considerations are prominent. The reliance on specific search terms for data collection presents a risk of excluding relevant content not captured by these terms, highlighting the need for iterative refinement of search queries and continuous validation of data. Additionally, the use of satellite articles from the Yale Humanitarian Research Lab, while reputable, might introduce inherent biases or focus areas that could skew the framing or interpretation of results. Furthermore, extracting messages and analyzing them outside their original context can sometimes lead to misinterpretations or a loss of nuanced meanings, underscoring the complexity of understanding and presenting data accurately.

VI. 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.

VII. Acknowledgments

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

References

[1] F. P. R. Institute, "Understanding russia's invasion of ukraine," https://www.fpri.org/, 2022, accessed: 2023-07-02.

[2] N. W. College, "Types of intelligence collection - intelligence studies," https://usnwc.libguides.com/intelligence/studies, 2023, accessed: 2023-07-02.

[3] D. Tsovaltzi, R. Judele, T. Puhl, and A. Weinberger, "Leveraging social networking sites for knowledge co-construction: Positive effects of argumentation structure, but premature knowledge consolidation after individual preparation," *Learning and Instruction*, vol. 52, pp. 161–179, 2017.

[4] J. S. Fu, "Leveraging social network analysis for research on journalism in the information age," *Journal of Communication*, vol. 66, pp. 299–313, 2016.

[5] C. P. Research, "Telegram becomes a digital forefront in the conflict - news feeds from fighting zones," https://blog.checkpoint.com/, 2023, accessed: 2023-07-02.

[6] B. Christensen and A. Khalil, "Reporting conflict from afar: Journalists, social media, communication technologies, and war," *Journalism Practice*, pp. 1–19, 2021.

[7] Q. Ge, M. Hao, F. Ding *et al.*, "Modelling armed conflict risk under climate change with machine learning and time-series data," *Nat Commun*, vol. 13, p. 2839, 2022.

[8] M. Colaresi and Z. Mahmood, "Do the robot: Lessons from machine learning to improve conflict forecasting," *Journal of Peace Research*, vol. 54, no. 2, pp. 193–214, 2017. [Online]. Available: http://www.jstor.org/stable/44511206

[9] K. Yang, T. Zhang, H. Alhuzali, and S. Ananiadou, "Cluster-level contrastive learning for emotion recognition in conversations," *arXiv preprint arXiv:2302.03508*, 2023. [Online]. Available: https://arxiv.org/abs/2302.03508

[10] T. Zeitzoff, "How social media is changing conflict," *The Journal of Conflict Resolution*, vol. 61, no. 9, pp. 1970–1991, 2017. [Online]. Available: http://www.jstor.org/stable/26363973

[11] I. Salehyan, "Best practices in the collection of conflict data," *Journal of Peace Research*, vol. 52, no. 1, pp. 105–109, 2015. [Online]. Available: http://www.jstor.org/stable/24557521

[12] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, no. 1. Citeseer, 2003, pp. 29–48.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[14] K. Proctor, "Social media and conflict: Understanding risks and resilience - research findings from ethiopia, iraq, myanmar, and nigeria," 7 2021. [Online]. Available: https://www.mercycorps.org/research-resources/analyzing-responding-social-media-conflict

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[16] K. S. Gleditsch, N. W. Metternich, and A. Ruggeri, "Data and progress in peace and conflict research," *Journal of Peace Research*, vol. 51, no. 2, pp. 301–314, 2014. [Online]. Available: http://www.jstor.org/stable/24557423

[17] Empirical Studies of Conflict, "Data," 2023. [Online]. Available: https://esoc.princeton.edu/data

[18] P. R. I. Oslo, "Data challenges in conflict research," https://blogs.prio.org, 2023.

[19] I. S. Association, "2013 annual meeting of the international studies association," Conference Proceedings, 2013.

[20] S. Dubberley, A. Koenig, and D. Murray, *Digital witness: using open source information for human rights investigation, documentation, and accountability*. Oxford University Press, USA, 2020.

[21] M. Lonkila, L. Shpakovskaya, and P. Torchinsky, "Digital activism in russia: The evolution and forms of online participation in an authoritarian state," *The Palgrave handbook of digital Russia studies*, pp. 135–153, 2021.

[22] A. International, "How is social media transforming human rights monitoring?" https://www.amnestyusa.org, 2023.

[23] "Cornell ConvoKit: A Collection of Conversations from Wikipedia Talk Pages," https://convokit.cornell.edu/documentation/awry.html, Cornell University, 2023, accessed: July 11, 2023.

[24] Y. H. R. Lab, "Conflict observatory," 2023, accessed: 2023-07-10. [Online]. Available: https://medicine.yale.edu/lab/khoshnood/projects/conflict-observatory/

[25] "Lyzem - Privacy Friendly Search Engine," https://lyzem.com/, Lyzem, 2023, accessed: July 11, 2023.

[26] FacebookAI, "roberta-base," https://huggingface.co/FacebookAI/roberta-base, 2019.

[27] R. Rehurek, "Gensim: Topic modelling for humans," https://radimrehurek.com/gensim/models/coherencemodel.html, 2022, accessed: 2023-06-13.
[28] P. Nemkova, S. Ubani, S. O. Polat, N. Kim, and R. D. Nielsen, "Detecting human rights violations on social media during russia-ukraine war," 2023.