TweetIntent@Crisis: A Dataset Revealing Narratives of Both Sides in the Russia-Ukraine Crisis

Lin Ai, Sameer Gupta, Shreya Oak, Zheng Hui, Zizhou Liu, Julia Hirschberg

Columbia University Department of Computer Science
lin.ai@cs.columbia.edu, {sg4021, so2667, zh2483, zl2889}@columbia.edu, julia@cs.columbia.edu

Abstract

This paper introduces **TweetIntent@Crisis**, a novel Twitter dataset centered on the Russia-Ukraine crisis. Comprising over 17K tweets from government-affiliated accounts of both nations, the dataset is meticulously annotated to identify underlying intents and detailed intent-related information. Our analysis demonstrates the dataset's capability in revealing fine-grained intents and nuanced narratives within the tweets from both parties involved in the crisis. We aim for **TweetIntent@Crisis** to provide the research community with a valuable tool for understanding and analyzing granular media narratives and their impact in this geopolitical conflict.

Introduction

The tensions between Russia and Ukraine, escalating over decades, reached a new peak in February 2022 following Russia's recognition of two Ukrainian breakaway regions (Hernandez 2022). As the conflict persists, a secondary battleground has unfolded in the digital realm. This new front is characterized by the strategic use of social media to rally support for both sides and plays a significant role in the broader context of information warfare.

In this paper, we aim to dissect the narratives from both Russian and Ukrainian perspectives regarding the crisis. We introduce **TweetIntent@Crisis**, a Twitter dataset featuring tweets from Russian and Ukrainian government-affiliated accounts. This dataset spans the year from February 2022 to February 2023, providing a comprehensive view of the crisis's evolution. It includes **17K cleaned source tweets** annotated with underlying intent and detailed intent-related information. Over 3K of these tweets are human-annotated with a meticulously designed annotation schema and process. The remaining 14K are machine-annotated using a pipeline incorporating multiple fine-tuned GPT-3.5-Turbo models.

We demonstrate that **TweetIntent@Crisis** is adept at revealing nuanced narratives and intricate intents behind the tweets. Our goal is for this dataset to assist the research community in delving deeper into the dynamics of how media narratives are constructed and disseminated on social media, and to understand their pivotal role in shaping the discourse of this crisis.

Related Work

Related Datasets

Since the start of the Russia-Ukraine crisis, many researchers have sought to better understand the conflict by collecting social media datasets on the topic. Some utilize the Twitter Streaming API with specific hashtags like #russia and #ukraine to conduct large-scale data collection (Haq et al. 2022; Smart et al. 2022; Pohl et al. 2023). Others employ the Twitter Historic Search API in combination with hashtags (Caprolu, Sadighian, and Di Pietro 2023; Park et al. 2022). Some studies integrate both methods for more flexible topic selection and content control (Chen and Ferrara 2023; Shevtsov et al. 2022). In addition to these datasets, Pohl et al. (2023) provide a comprehensive survey of the current social media datasets on this issue.

However, a common limitation of these collection processes is the presence of noisy, less informative data and the lack of source credibility. To address this, our study focuses on ensuring information quality and source reliability by specifically collecting tweets from accounts associated with Russian and Ukrainian government media and organizations. This approach guarantees high-quality data that not only offers substantial information but also reflects the official stances of the countries to a considerable degree.

Propaganda And Intent

The study of propaganda in social media has gained substantial attention, particularly given the rise of social media as a prime platform for information warfare, enabling the mass dissemination of information. The escalation of the Russo-Ukrainian conflict in 2022 has spurred numerous research efforts focused on analyzing social media propaganda campaigns, notably those originating from Russia. Golovchenko (2022) investigates the censorship of Ukrainian content on Russian social media platforms. Geissler et al. (2023) examine the proliferation of pro-Russian sentiment on social media and the role of *bots* in amplifying these messages. Soares, Gruzd, and Mai (2023) examine the characteristics of online users inclined to believe in pro-Kremlin disinformation narratives.

In addition to research concentrating on event-specific and demographic analyses, there is a notable body of work examining propaganda techniques in textual content. Researchers have developed various lists categorizing these techniques (Torok 2015). For example, Miller (1939) introduces a foundational classification of propaganda, comprising seven techniques. Habernal, Pauli, and Gurevych (2018) construct a corpus of 1.3k arguments, annotated with five fallacies directly associated with propaganda techniques. Da San Martino et al. (2019) build upon previous studies, identifying 18 propaganda techniques and developing a corpus of news articles for the detection and classification of these techniques. However, while these studies establish frameworks for detecting techniques in messages, there is still a need for further research to understand the intents behind the use of these techniques in individual messages. Some studies suggest a vague concept of intent (Ai et al. 2021; Gabriel et al. 2021), or attempt to model individual intent behind spreading fake news (Guo et al. 2023; Zhou et al. 2022). Our work seeks to further explore this issue. Additionally, while many studies focus primarily on the perspective of the Russian party, we are particularly interested in understanding the nuanced narratives from both sides, Russia and Ukraine, involved in the conflict. This more balanced approach allows us to gain a more comprehensive view of the entire situation, encompassing the diverse viewpoints and strategies employed by each party.

Data Collection

Government-Affiliated Tweets Collection

To construct our corpus, we collected tweets over the course of one year, from February 1, 2022 and to February 28, 2023, to capture the evolution of the crisis. This method aimed to provide a detailed understanding of how the crisis unfolded. To guarantee the reliability of our data, we initiated our collection with a manually-selected set of seed accounts labeled ”state-affiliated media” by Twitter, as shown in Table 1. This approach ensured that we only gathered tweets from sources verified by Twitter as government-affiliated. Following the selection of these seed accounts, we recursively extended our search to include their followers, focusing on accounts categorized as ”state-affiliated media”, ”government official”, and ”government-funded media”. We continued this expansion until no additional accounts were found. Currently, our collection consists of 67 Russian and 12 Ukrainian accounts, detailed in Table 10 in the Appendix. The significant difference in the number of accounts is due to the larger presence of Russian accounts identified as government-affiliated on Twitter compared to Ukrainian accounts during our data collection period. Following this, we utilized the Twitter Historic Search API to retrieve all English-only content posted by these accounts during the specified time-frame using the Twitter Search API.

	Seed Accounts
Russian Accounts	@RT_com, @SputnikInt @redfishstream, @tassagency_en @Ruptly
Ukrainian Accounts	@United24media

Table 1: Seed Set of State-Affiliated Accounts

Topic Modeling

To ensure the relevance and quality of our data, we applied Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) for topic modelling and filtered out irrelevant tweets, ensuring that our data is focused solely on the Russia-Ukraine crisis.

We preprocessed the tweets through tokenization, removing stopwords and URLs, and lemmatization. Recognizing the distinct nature of tweets from Russian and Ukrainian accounts in terms of topics and audience, we fine-tuned the hyperparameters — specifically, learning decay and the number of topics — separately for each set of tweets. The optimization of these parameters was conducted using a grid search method, with topic coherence (Röder, Both, and Hinneburg 2015) as the evaluation metric. The resulting top salient terms for each topic identified by the LDA topic modeling algorithm in tweets from Russian and Ukrainian accounts are displayed in Tables 11 and 12 in the Appendix. Subsequently, we focused on tweets associated with specific keywords of interest, such as ”war”, ”invasion”, ”armed”, ”kill”, ”civilian”, ”attack”, ”defend”, ”enemy”, ”hero”, and ”victory”. We only included tweets in topics where at least three keywords appeared among the top 20 salient terms.

After completing the data cleaning and LDA topic modeling process, our final dataset consists of **17,854 cleaned source tweets**. These include only original tweets and exclude retweets or replies.

Data Annotation

Annotation Schema

In the Related Work section, we discuss how propaganda techniques have been extensively researched, yet the specific intent behind each message remains less explored. These intents, while not always strictly classified as propaganda, can significantly influence the audience’s perception and sway opinions towards certain events or entities. Our analysis of the 18 techniques identified by Da San Martino et al. (2019) reveals that the application of these techniques typically aligns with one of two primary intents, as defined below:

> **Call To Action (*CTA*):** This intent assesses whether a particular message is designed to prompt its target audience to take specific action or address an issue. Such content is intended to compel the audience to undertake a particular task or respond immediately, often framed as an instruction or directive. The intended audience for these messages can be any entities mentioned in the message or the readers themselves.
>
> **Discredit Entity (*DE*):** This intent determines whether a message aims to harm the reputation, credibility, or competence of an individual, organization, or nation-state. It often features a negative tone, possibly employing loaded language, slurs, or unfavorable comparisons to negatively perceived semantic categories.

Table 2 presents a detailed mapping of the 18 propaganda techniques identified by Da San Martino et al. (2019) in relation to the two primary intents. It categorizes each technique based on their typical usage in conveying these intents.

	Common Techniques
CTA	flag-waving, slogans, appeal to fear, repetition, appeal to authority, dictatorship, black-and-white fallacy, bandwagon, reductio ad hitlerum, thought-terminting cliché
DE	loaded language, name calling, obfuscation, exaggeration/minimizat, doubt, causal oversimplification, whataboutism, staw man, red herring, thought-terminting cliché

Table 2: Mapping of Techniques to Intents

In addition, we not only focus on identifying but also on characterizing detected intents by pinpointing the involved subjects and the specific information conveyed. Specifically, for ***CTA***, we aim to identify the following *fields*:

Called Subjects The intended audience for the calls.
Called Actions The particular actions being urged.

Regarding ***DE***, our goal is to identify the following *fields*:

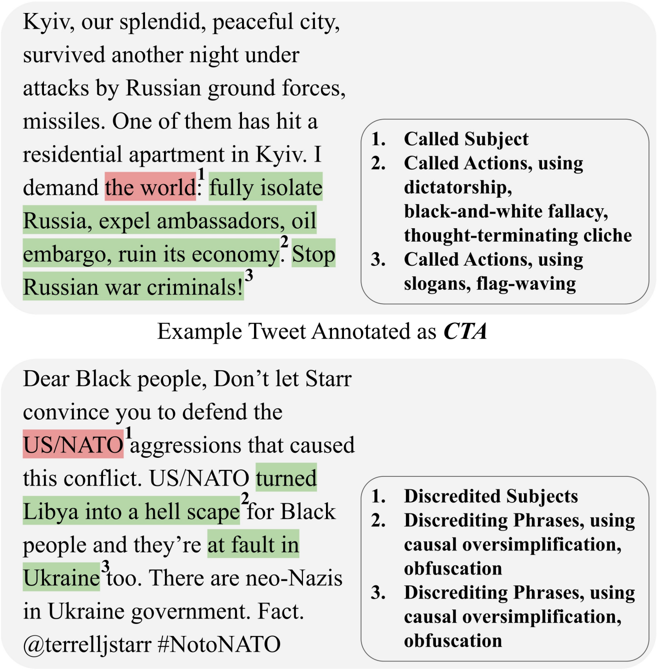
Discredited Subjects The entities that are being undermined or criticized.
Discrediting Phrases The specific phrases used to discredit these entities.

Figure 1 provides examples of tweets illustrating these two intents, along with the detailed information we seek to pinpoint. The following sections elaborate on our data annotation process, which involves annotating both the intents and their associated details through methods including human annotation and larger-scale machine annotation using large language models (LLM).

Our research concentrates primarily on discerning the intent behind tweets, rather than identifying propaganda techniques or gauging the level of propaganda. Our goal is not to learn about propaganda itself; rather, we aim to comprehensively understand the narratives propagated by different governments. It is crucial to note that tweets annotated or labeled as containing either ***CTA*** or ***DE*** should not be taken or classified as propaganda tweets. These labels serve to identify the underlying intent or narrative strategy, and do not necessarily imply the presence of propaganda.

Human Annotation

Pre-Annotation Due to the high costs and time constraints associated with human annotation, we employed GPT-4 (OpenAI 2023) for an initial round of pre-annotation on the tweets. This preliminary step was crucial for filtering out tweets without potential intents, thus streamlining the process for more focused human annotation. Specifically, for each tweet, we used GPT-4 to assess whether it called for any actions (***CTA***) or discredited any entities (***DE***). The specific queries applied in this pre-annotation phase are detailed in Table 13 in the Appendix. After the pre-annotation process, we selected **5,000** tweets that GPT-4 labeled as containing either ***CTA*** or ***DE*** intent for human annotation.


Kyiv, our splendid, peaceful city, survived another night under attacks by Russian ground forces, missiles. One of them has hit a residential apartment in Kyiv. I demand the world[1]: fully isolate Russia, expel ambassadors, oil embargo, ruin its economy[2]. Stop Russian war criminals![3]

1. **Called Subject**
2. **Called Actions, using dictatorship, black-and-white fallacy, thought-terminating cliche**
3. **Called Actions, using slogans, flag-waving**

Example Tweet Annotated as ***CTA***

Dear Black people, Don't let Starr convince you to defend the US/NATO[1] aggressions that caused this conflict. US/NATO turned Libya into a hell scape[2] for Black people and they're at fault in Ukraine[3] too. There are neo-Nazis in Ukraine government. Fact. @terrelljstarr #NotoNATO

1. **Discredited Subjects**
2. **Discrediting Phrases, using causal oversimplification, obfuscation**
3. **Discrediting Phrases, using causal oversimplification, obfuscation**

Example Tweet Annotated as ***DE***

Figure 1: Example Tweets Annotated With Intents

Annotation Setup We collaborated with four well-trained English-speaking annotators for the task of annotation. Each tweet was assigned to three of these annotators to ensure annotation quality. For each tweet, we requested the annotators to execute the following steps:

1. Determine the presence of either ***CTA*** or ***DE***. This involves answering two distinct binary questions:

   ***CTA*:** Is this tweet calling for any actions?
   ***DE*:** Are there any entities discredited in the tweet?

2. If either ***CTA*** or ***DE*** is identified, answer two interrogative questions about these intents. These questions aim to pinpoint the specific *fields*, including subjects involved and the specific information conveyed. Locate the answers within the tweet and highlight them as text spans:

   a. If ***CTA*** presents:
      Called Subjects: Who are being called upon?
      Called Actions: What actions are being called for?

   b. If ***DE*** presents:
      Discredited Subjects: Who are being discredited?
      Discrediting Phrases: How are they discredited?

A single tweet may exhibit one, two, or neither of the intents. For each interrogative question, annotators were allowed to highlight multiple text spans as answers within the tweet.

We developed a custom data annotation platform utilizing Label Studio[1], an open-source tool designed for creating data annotation interfaces and back-ends. This platform enabled annotators to respond to multiple-choice questions and

[1] https://labelstud.io/

to highlight specific text spans within tweets for answering the interrogative localization questions. A screenshot showcasing the user interface of our platform is shown in Figure 10 in the Appendix.

We provided detailed instructions and definitions of intent terms through dropdown menus in our annotation platform, ensuring that annotators could easily access and refer to these guidelines throughout the annotation process. For the task of highlighting text spans in response to interrogative questions, annotators were instructed to select the most concise spans possible. This involved avoiding unnecessary elements like stop words and punctuation marks.

Importantly, we emphasized the need for neutrality, instructing annotators to avoid personal political biases during the process. Fortunately, while the tweets pertain to the politically sensitive topic of the Russo-Ukraine conflict, our task centers on identifying intent, subjects, and information based on lexical and linguistic features, without the influence of personal political views. The involvement of three annotators for each tweet serves not only to obtain thorough analysis but also helps in cross-verifying the annotation quality and mitigating potential biases.

Post-Processing Despite comprehensive instructions, inevitable errors in annotation did occur, particularly during the initial stages. These errors were primarily due to confusion about using the annotation platform. Common issues included incorrect logging of answers and unintentional blank annotations. We conducted thorough inspections and utilized scripts to identify and remove these erroneous annotations. After this cleaning process, from the original 5,000 annotated tweets, we obtained a dataset of **3,691** tweets with valid annotations, as detailed in Table 3.

Since each tweet was annotated by three annotators, post-processing was essential to consolidate these multiple annotations into a single record for each tweet. For binary questions concerning intent identification, we adopted a majority voting approach to determine the final intent annotations for each tweet. In the case of span localization for interrogative questions, we considered only those annotations that aligned with the binary majority vote results. From these, we selected the longest common substrings as the final span annotations for each tweet. After the automated post-processing stage, we conducted an additional manual check to verify that the annotations were accurately recorded and merged.

Table 3 presents the statistics of our annotated dataset. Following the post-processing stage, we identified 93 tweets annotated with the presence of ***CTA*** and 411 tweets labeled with ***DE***. A notable challenge we faced was the limited number of positive samples, which posed difficulties for further analysis and model training. This limitation also underscores the complexity of this task for current LLMs, such as GPT-4. Despite pre-selecting tweets using GPT-4, only a small fraction was identified by human annotators as containing the specified intents. In response to this challenge, we explored methods to enhance the capability of LLMs in performing this task more effectively, which will be detailed in subsequent sections.

	All Annotations	**Valid Annotations**
#Tweets	5,000	3,691
	CTA	***DE***
#Tweets	93	411

Table 3: Annotation Statistics

Annotator Agreements We examined inter-annotator agreement to further validate the quality of this annotation. For binary classification regarding intent identification, we used Fleiss Kappa to measure the inter-annotator agreements. As listed in Table 4, the Kappa score is 0.863 for ***CTA*** identification, and 0.804 for ***DE*** identification, both indicating almost perfect agreements.

To evaluate annotator agreement on span localization for interrogative questions, we measured the overlap between each pair of annotations. For each pair of text spans s_i and s_j, we define the pairwise retrieved scores as follows:

$$\begin{aligned} r_{ij} &= len(LCS(s_i, s_j))/len(s_i) \\ r_{ji} &= len(LCS(s_i, s_j))/len(s_j) \end{aligned} \tag{1}$$

Here, LCS denotes the longest common substring. Assuming annotation 1 (ann_1) contains n text spans and annotation 2 (ann_2) contains m text spans, the retrieved score for each s_i in ann_1 is the highest score between it and all s_j in ann_2 for $j \in [1, m]$. The text span agreement between two annotations is then defined as a pairwise F1 score, computed as:

$$\begin{aligned} r_{ann_1} &= \frac{\sum_{i=1}^{n} max_{j\in[1,m]}(r_{ij})}{n} \\ r_{ann_2} &= \frac{\sum_{j=1}^{m} max_{i\in[1,n]}(r_{ji})}{m} \\ F1 &= \frac{2 * r_{ann_1} * r_{ann_2}}{r_{ann_1} + r_{ann_2}} \end{aligned} \tag{2}$$

The agreement among three annotations for each tweet is then the average of all pairwise F1 scores. The overall annotation agreement for a specific *field* (such as ***Called Subjects*** or ***Called Actions***) is the average of all tweets' agreements in that *field*.

As shown in Table 4, the average F1 agreement scores for ***Called Subjects*** and ***Called Actions*** are 0.988, and for ***Discredited Subjects*** and ***Discrediting Phrases***, are 0.936 and 0.920, respectively. These scores suggest almost perfect agreement. However, to account for the majority of tweets labeled without any intents (thus with empty span annotations), we also examined the agreement scores on positive samples. For tweets labeled with ***CTA***, the F1 scores for ***Called Subjects*** and ***Called Actions*** are 0.683 and 0.717, respectively. For tweets with ***DE***, the scores for ***Discredited Subjects*** and ***Discrediting Phrases*** are 0.787 and 0.644, respectively. These substantial agreement scores among annotators on span localization annotations affirm the quality of the annotations achieved.

Binary Classification	***CTA***	***DE***
Fleiss Kappa	0.863	0.804
Span Localization (*CTA*)	***Called Subjects* F1**	***Called Actions* F1**
All Valid Annotations	0.988	0.988
***CTA* Annotations**	0.683	0.717
Span Localization (*DE*)	***Discredited Subjects* F1**	***Discrediting Phrases* F1**
All Valid Annotations	0.936	0.920
***DE* Annotations**	0.787	0.644

Table 4: Inter-Annotator Agreements

Machine Annotation

In the previous sections, we explain that the human-annotated dataset comprises a limited number of positive samples. This limitation presents significant challenges for the further analysis. It also highlights the complexity of the task and the difficulties faced by current state-of-the-art (SOTA) LLMs. Despite selecting 5000 tweets for human annotation using GPT-4, only a small proportion is recognized by human annotators as having the specified intents. However, the gold labels obtained from human annotation enable us to improve the performance of LLMs in this task. This enhancement, in turn, allows us to conduct a more extensive machine-annotation on our entire dataset of 17K tweets.

Intent Binary Classification For intent binary classification, we explored three approaches:

1. **Zero-Shot GPT-4-Turbo:** We utilized the gpt-4-1106-preview model from OpenAI for identifying ***CTA*** and ***DE***. Queries used are listed in Table 14 in the Appendix. This approach differed from the pre-annotation stage, as we input a tailored system message directing the model to act as a social media content moderator. The output was constrained to JSON format to facilitate subsequent processing.
2. **GPT-4-Turbo with In-Context Learning (ICL):** Here, the GPT-4-Turbo model received both a positive and a negative example for classification, along with the same queries and system message as in the zero-shot setting. These examples were manually chosen from our human-annotated dataset.
3. **Fine-Tuned (FT) GPT-3.5-Turbo:** Additionally, we fine-tuned the GPT-3.5-Turbo model using our human-annotated dataset, maintaining a 3/7 ratio for testing and training. The queries for this model were identical to those used in the zero-shot setting. Details on fine-tuning are provided in the Appendix.

Table 5 presents the performance of these models on the human-annotated dataset, highlighting both the overall classification F1 score and the true class (that intent exists) F1 score. For the fine-tuned GPT-3.5-Turbo, we report results from the test set, which comprises 30% of all human-annotated data, as well as from the entire human-annotated dataset. The test set results are our primary focus for comparisons, while the full dataset results serve as supplementary reference. The fine-tuned GPT-3.5-Turbo demonstrates superior performance in both ***CTA*** and ***DE*** classifications. Specifically, it surpasses GPT-4-Turbo by 8.79% in overall F1 score and exceeds GPT-4-Turbo with ICL by 13.79% in ***CTA*** classification. In ***DE*** classification, it outperforms GPT-4-Turbo by 75.93% and GPT-4-Turbo with ICL by 41.79% in overall F1 score.

It is important to note that without fine-tuning, the models exhibit significantly lower performance in true class classification. This is primarily due to the imbalance in our dataset, in which a majority of the tweets are negative samples. Fine-tuning effectively addresses this issue. Specifically, in ***CTA*** classification, the fine-tuned GPT-3.5-Turbo shows a remarkable improvement in the true class F1 score, enhancing it by 133.33% compared to GPT-4-Turbo, and by 185.19% compared to GPT-4-Turbo with ICL. Likewise, in ***DE*** classification, the fine-tuned model boosts the true class F1 score by 140.63% compared to GPT-4-Turbo, and by 102.63% compared to GPT-4-Turbo with ICL.

	CTA	***DE***
GPT-4-Turbo	0.91 | 0.33	0.54 | 0.32
GPT-4-Turbo+ICL	0.87 | 0.27	0.67 | 0.38
FT GPT-3.5-Turbo		
Test Set	**0.99** | **0.77**	**0.95** | **0.77**
All Human-Annotated Data	0.99 | 0.84	0.97 | 0.84

Table 5: Performance (Overall F1 | True Class F1) in Intent Binary Classification of Different LLM Models

Interrogative Span Localization Similarly, for the interrogative span localization task, we employed three approaches: **Zero-Shot GPT-4-Turbo**, **GPT-4-Turbo with ICL**, and **FT GPT-3.5-Turbo**. The system message and queries utilized in these approaches were consistent across all three and can be found detailed in Table 15 in the Appendix. Additionally, we employed the function-calling feature of the OpenAI GPT API to standardize and consolidate the output format. The specific functions defined and utilized for this purpose are detailed in the Appendix. In the ICL approach, we provided one carefully selected example, annotated with the interrogative text spans, along with the identical system message and queries as those in the zero-shot approach.

Table 6 shows the performance of these approaches on our human-annotated datasets. We used two metrics: exact match (EM) F1 and overlap F1, following the methodology of Li et al. (2022) used to evaluate model performance in multi-span span selection tasks. Specifically for the fine-tuned GPT-3.5-Turbo, we present results from the test set, which constitutes 30% of the entire human-annotated dataset. These results are our primary benchmark for comparing performance. Additionally, we include results from the full human-annotated datasets as a supplementary reference.

Overall, the fine-tuned GPT-3.5-Turbo outperforms significantly across metrics. In ***CTA***, it exceeds GPT-4-Turbo with ICL by 11.67% in ***Called Subjects*** EM F1 and 2.67% in overlap F1, and by 8.11% in EM F1 and 1.23% in overlap F1 for ***Called Actions***. For ***DE***, it surpasses GPT-4-Turbo with ICL by 12.28% in ***Discredited Subjects*** EM F1 and 6.85% in overlap F1, and by 58.33% in ***Discrediting Phrases*** EM F1 and 7.25% in overlap F1.

CTA	***Called Subjects***	***Called Actions***
GPT-4-Turbo	0.49 | 0.62	0.29 | 0.76
GPT-4-Turbo+ICL	0.60 | 0.75	0.37 | 0.81
FT GPT-3.5-Turbo		
Test Set	**0.67** | **0.77**	**0.40** | **0.82**
All Human-Annotated Data	0.81 | 0.85	0.65 | 0.88
DE	***Discredited Subjects***	***Discrediting Phrases***
GPT-4-Turbo	0.55 | 0.73	0.26 | 0.68
GPT-4-Turbo+ICL	0.57 | 0.73	0.24 | 0.69
FT GPT-3.5-Turbo		
Test Set	**0.64** | **0.78**	**0.38** | **0.74**
All Human-Annotated Data	0.77 | 0.86	0.59 | 0.85

Table 6: Performance (EM F1 | Overlap F1) in Interrogative Span Localization of Different LLM Models

Machine Annotation Pipeline Subsequently, we implemented a two-step pipeline for machine-annotating our dataset's unannotated portion, circumventing the high costs of manual annotation. First, we used a fine-tuned GPT-3.5-Turbo model for binary intent classification of tweets. Subsequently, we annotated tweets with identified intents for interrogative span localization, again employing a GPT-3.5-Turbo model fine-tuned for this purpose.

Table 7 provides statistical details of **Tweet-Intent@Crisis**. The dataset comprises a total of 17,854 cleaned source tweets. Out of these, 3,691 tweets are human-annotated, as described in the previous section. The remaining 14,163 tweets are machine-annotated using our machine annotation pipeline, with 307 tweets are labeled as containing ***CTA***, and 767 are identified as containing ***DE***.

Full dataset #Tweets	17,854	
	Human Annotated	**Machine Annotated**
#Tweets	3,691	14,163
#CTA Tweets	93	307
#CTA Text Spans	196	537
#DE Tweets	411	767
#DE Text Spans	1,292	2,447

Table 7: Dataset Statistics

Content Analysis

Aggregated Content Analysis

We conducted a further content analysis of **Tweet-Intent@Crisis** to gain insights into the content and narratives shared by both sides in the Russia-Ukraine conflict. As shown in Table 8, of the 17K tweets in our dataset, approximately 14K are from Russian accounts, while around 3K are from Ukrainian accounts.

Hashtags To understand the focus of the discussions in the tweets, we first examined the hashtags used. Figures 2 and 3 display the top 20 hashtags frequently used by Russian and Ukrainian accounts, respectively. We observe that the hashtags are closely related to the Russia-Ukraine crisis, confirming the efficacy of our topic modeling process used during data collection. Ukrainian accounts frequently use hashtags like #standwithukraine, #stoprussianagression, and #armukrainenow, indicating a call for support and conveying urgency. In contrast, Russian tweets include various tags mentioning relevant parties or entities, such as #us, #nato, and #putin, suggesting a tendency towards sharing information or theories, or other intents, rather than solely focusing on rallying support.

	Russian Accounts	**Ukrainian Accounts**
#Tweets	14,350	3,504
#*CTA* Tweets	105	295
#*DE* Tweets	682	496

Table 8: #Tweets from Russian and Ukrainian Accounts

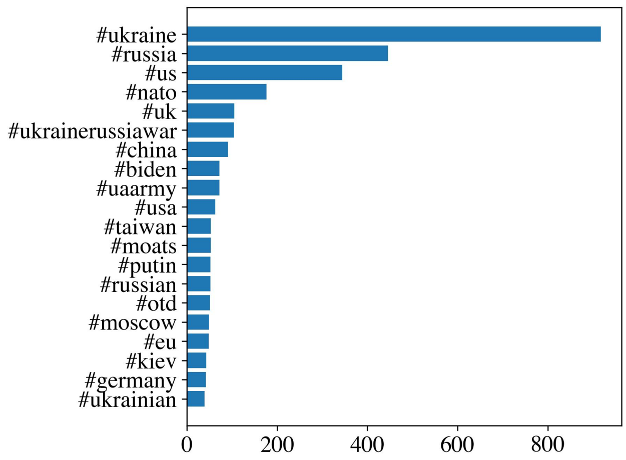


Figure 2: Top 20 Hashtags Used in Tweets from Russian Accounts

Domains We also examined the URLs shared in the tweets to better understand the information sources used to distribute, support, and validate their content. Our focus was primarily on domains linked to news agencies and media sources. We excluded social media domains such as Facebook, Instagram, and Telegram, as well as video sharing platforms like YouTube. Our findings show that tweets from Russian accounts frequently cite content from Russia's state-owned news agencies and media sources. This is partly because our data collection includes tweets directly from the official accounts of these agencies. Table 9 presents the top 5 sources quoted in tweets from Russian accounts. In contrast, while Ukrainian accounts are also government-affiliated and include media and organization accounts, we do not observe

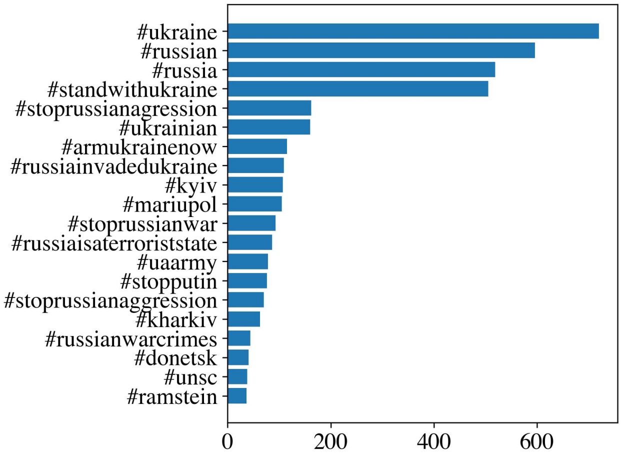


Figure 3: Top 20 Hashtags Used in Tweets from Ukrainian Accounts

a similar pattern of frequently quoted common sources.

Top 5 Sources
RT
Sputnik News
TASS
Russia Beyond
Ruptly

Table 9: Top 5 Media Sources Most Frequently Quoted in Tweets from **Russian** Accounts

Granular Narrative Analysis

TweetIntent@Crisis, annotated with specific intents, allows us to conduct a detailed and granular analysis of the narratives from both parties. Our focus is on analyzing tweets identified with ***CTA*** and ***DE*** intents. This approach gives us a unique perspective to understand the underlying intent in these narratives, including the audience targeted and the specific information conveyed.

CTA We combined human-annotated and machine-annotated tweets, resulting in a total of 105 tweets from Russian accounts and 295 tweets from Ukrainian accounts identified with ***CTA***, as detailed in Table 8. This observation is consistent with our earlier findings indicating that Ukrainian accounts are more focused on rallying support compared to Russian accounts. This is evident from the higher proportion of Ukrainian tweets labeled as ***CTA***, despite their overall smaller tweet count.

We examined the ***Called Subjects*** and ***Called Actions*** spans within the annotated tweets. Figures 4 and 5 present Word Clouds representing these spans in tweets from Russian and Ukrainian accounts, respectively. These Word Clouds reveal that Russian accounts predominantly address a global audience, urging collective efforts to stop the war, with peace activists and politicians being other frequent targets. Ukrainian accounts, while also appealing to the world and the international community for action against the war, differ in their specific calls. They frequently advocate for sanctions against Russia and request weapon support. Additionally, Ukrainian tweets often target UN members and their partners, highlighting a broader and more diverse range of addressed audiences compared to Russian accounts.

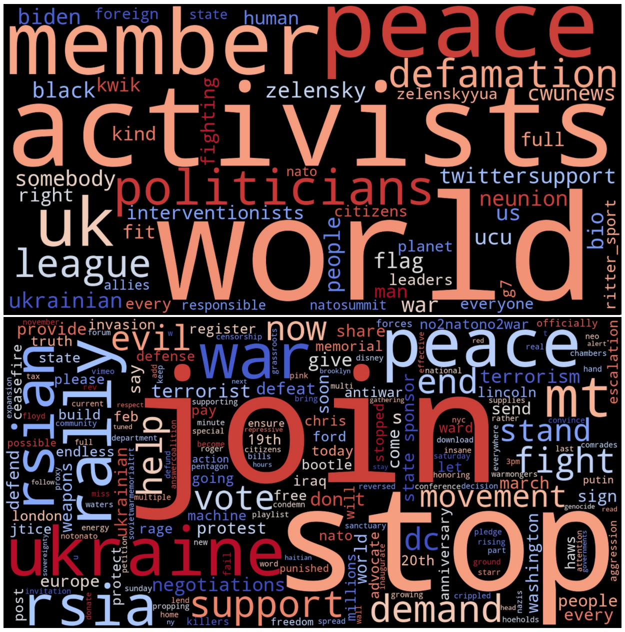


Figure 4: Word Clouds of the Most Frequent ***Called Subjects*** (Top) and ***Called Actions*** (Bottom) Spans in Tweets from **Russian** Accounts

DE Similarly, we combined both human-annotated and machine-annotated tweets, yielding a total of 682 tweets from Russian accounts and 496 tweets from Ukrainian accounts identified with ***DE***, as shown in Table 8. We analyzed the ***Discredited Subjects*** and ***Discrediting Phrases*** spans by examining their respective Word Clouds, as illustrated in Figures 6 and 7. These figures reveal that Russian accounts frequently target and discredit entities such as Ukraine, NATO, and Western governments. The narrative seems to focus on blaming these entities for the war, civilian casualties, and labeling them as Nazis and war criminals. In contrast, Ukrainian accounts primarily discredit Russia, attributing to it the acts of invasion, and the killing of civilians and children. Notably, Russia appears as the sole major subject of discredit in the Ukrainian narrative.

It is important to note that these nuanced narratives are uncovered thanks to our annotated dataset. For comparison, examining the aggregated Word Clouds from all tweets from Russian and Ukrainian accounts, as shown in Figure 8, reveals only a set of frequently occurring terms with similar patterns. For example, when we encounter high-frequency named entities, it is not clear whether they are the intended audience of the messages or the subjects being discredited.

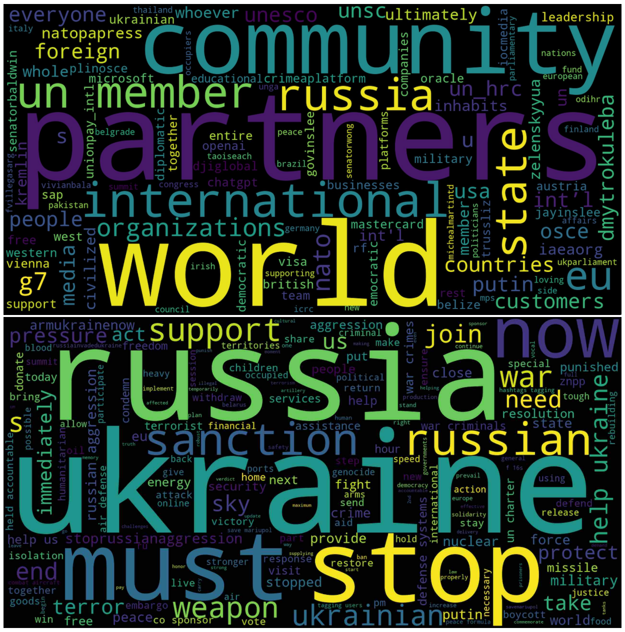


Figure 5: Word Clouds of the Most Frequent ***Called Subjects*** (Top) and ***Called Actions*** (Bottom) Spans in Tweets from **Ukrainian** Accounts

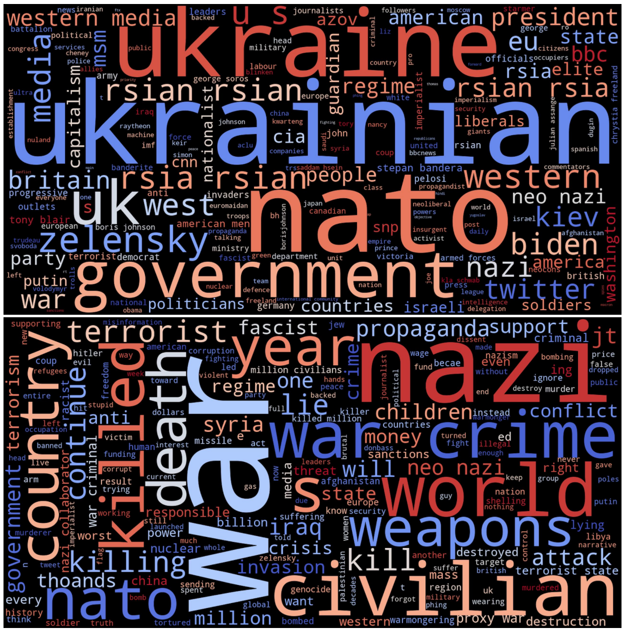


Figure 6: Word Clouds of the Most Frequent ***Discredited Subjects*** (Top) and ***Discrediting Phrases*** (bottom) Spans in Tweets from **Russian** Accounts

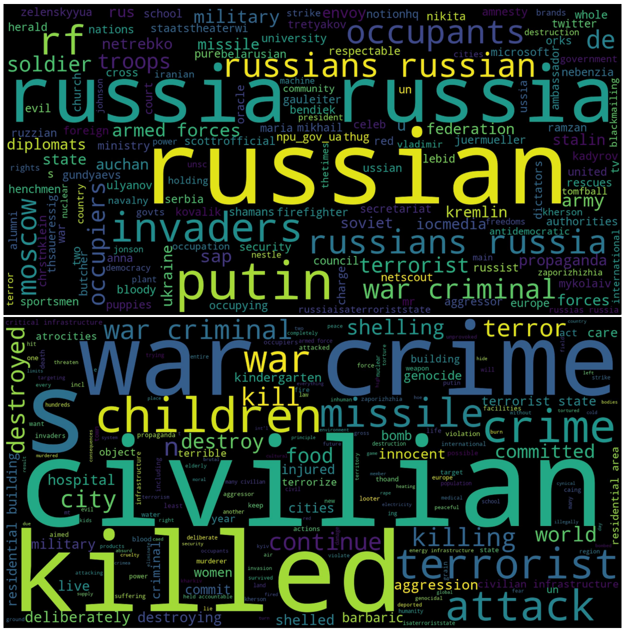


Figure 7: Word Clouds of the Most Frequent ***Discredited Subjects*** (Top) and ***Discrediting Phrases*** (bottom) Spans in Tweets from **Ukrainian** Accounts

This ambiguity was resolved by analyzing our annotated dataset, as demonstrated in the previous sections.

Additionally, while the crisis predominantly involves Russia and Ukraine, our analysis uncovers the involvement of other parties. For instance, Ukrainian accounts frequently appeal to their Western partners for support, who are simultaneously subject to significant discredit by Russian accounts. This reveals the positions and reactions of various global entities to the conflict. Although this information may seem apparent to those familiar with the context, our dataset and annotation schema provide an automated, effective way to discern and understand these dynamics.

Release and Access

TweetIntent@Crisis is available at the following URL:

https://doi.org/10.5281/zenodo.10499589

Due to the Twitter Terms of Service (ToS)[2], we are limited to publicly redistributing only certain information, such as Tweet IDs. Additionally, hydrated tweets can be shared under restricted conditions with a cap of 50,000 tweets. To adhere to these terms, we release a minimal set of hydrated tweets, specifically those annotated with ***CTA*** or ***DE***, encompassing both human-annotated and machine-annotated tweets. The remainder of the dataset is released with only Tweet IDs.

[2] https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases

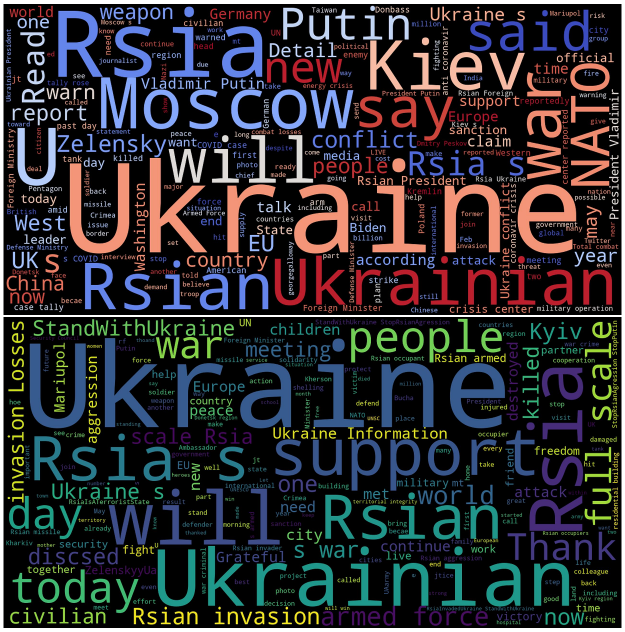


Figure 8: Word Clouds of the Most Frequent Terms in Tweets from **Russian** Accounts (Top) and **Ukrainian** Accounts (Bottom

Conclusions

We present **TweetIntent@Crisis**, a novel Twitter dataset comprising tweets from Russian and Ukrainian government-affiliated accounts. This dataset contains 17K source tweets, each annotated to identify the underlying intent and highlight detailed intent-related information. Of these, over 3K tweets are human-annotated, while the rest are machine-annotated through a pipeline involving several fine-tuned GPT-3.5-Turbo models.

Our analysis of the tweet content illuminates the intricate narratives crafted by both sides in the conflict. By examining hashtags and domains, we identify distinct focal points and sources of information for each party. Further, our scrutiny of tweets annotated with ***CTA*** and ***DE*** intents offers a more detailed view, exposing diverse targets, strategies, and narratives used by each side to appeal for support and discredit their opponents. Additionally, our study provides insights into the roles and responses of other global players in the crisis, enriching our understanding of this complex geopolitical situation.

Ethical Statement

Data Collection Our data collection process utilizes the Twitter Historic Search API, ensuring that all data gathered are from publicly available information at the time of collection. This approach is in compliance with the Twitter ToS.

Data Annotation The collection and annotation of this dataset have been IRB-approved by the Aptima Institutional Review Board for our DARPA SemaFor (Semantic Information Defender) Project with Kitware Inc. All annotators, including those involved in human annotation and post-annotation data verification, have participated voluntarily. They are fully informed about the task and any potential risks of harm related to their participation. Moreover, none of the annotators or researchers involved in this project are immediate parties in the conflict, specifically, they are neither Russian nor Ukrainian.

All annotators have acknowledged the importance of maintaining neutrality and agreed to avoid personal political biases during the annotation process. Although our task is largely objective, focusing mainly on the annotators' understanding of lexical and linguistic content rather than their personal political views, this measure ensures the integrity and impartiality of the dataset.

Data Release In adherence to Twitter ToS, we release only a minimal set of hydrated tweets, particularly those annotated with ***CTA*** or ***DE***. The rest of the dataset is shared solely via their Tweet IDs. As the dataset includes information beyond just Tweet IDs, access is restricted and available only upon approved request. Specifically, requests will only be approved for users affiliated with a research agency or institute, and the dataset is to be used strictly for research and non-commercial purposes.

Data Analysis Ethical considerations regarding user anonymity are crucial in our research. Tweet objects inherently contain information about users and their accounts. Users can opt to restrict access to their tweets via the API by either setting their accounts to private or by deleting their tweets. To address these concerns about user privacy, our paper focuses solely on presenting aggregated statistics and avoids disclosing individual user data. This approach is designed to respect user privacy while still providing valuable insights from the dataset.

Limitations

It is important to acknowledge the limitations of this dataset. First, it focuses exclusively on English-language tweets, potentially missing key aspects of the multilingual discourse on the topic. The data is collected using the Twitter Historic Search API and topic modeling techniques. While it spans from February 2022 to February 2023, it may not capture all relevant tweets, possibly omitting some aspects of the event. Moreover, there is a possibility that some tweets may have been deleted or made private, or are posted by users who have since either made their accounts private, had their accounts suspended, or deleted their accounts after our data collection period. This presents potential challenges for future researchers in accessing or utilizing these specific tweets in their studies.

Despite efforts to filter out sensitive content, the dataset might still include content raising privacy concerns, a common issue with user-generated content on social media. The dataset predominantly comprises tweets from Twitter labeled accounts, affecting the nature of the information collected. These limitations warrant careful consideration in interpreting the findings and in their broader generalization.

A portion of our dataset has been machine-annotated using fine-tuned GPT-3.5-Turbo models. Although these annotations are useful for aggregate analysis and model development, there may be instances of misclassification and errors due to the limitations of the models. Therefore, these annotations should not be regarded as definitive gold labels, and caution should be exercised in their future application.

Acknowledgments

This research was done with funding from the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0123. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

The authors gratefully acknowledge Kitware Inc. for their invaluable contribution to the human annotation process.

References

Ai, L.; Chen, R.; Gong, Z.; Guo, J.; Hooshmand, S.; Yang, Z.; and Hirschberg, J. 2021. Exploring New Methods for Identifying False Information and the Intent Behind It on Social Media: COVID-19 Tweets. In *ICWSM Workshops*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022.

Caprolu, M.; Sadighian, A.; and Di Pietro, R. 2023. Characterizing the 2022-Russo-Ukrainian conflict through the lenses of aspect-based sentiment analysis: dataset, methodology, and key findings. In *2023 32nd international conference on computer communications and networks (ICCCN)*, 1–10. IEEE.

Chen, E.; and Ferrara, E. 2023. Tweets in time of conflict: A public dataset tracking the twitter discourse on the war between Ukraine and Russia. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 1006–1013.

Da San Martino, G.; Yu, S.; Barrón-Cedeño, A.; Petrov, R.; and Nakov, P. 2019. Fine-Grained Analysis of Propaganda in News Articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, EMNLP-IJCNLP 2019. Hong Kong, China.

Gabriel, S.; Hallinan, S.; Sap, M.; Nguyen, P.; Roesner, F.; Choi, E.; and Choi, Y. 2021. Misinfo Reaction Frames: Reasoning about Readers' Reactions to News Headlines. *arXiv preprint arXiv:2104.08790*.

Geissler, D.; Bär, D.; Pröllochs, N.; and Feuerriegel, S. 2023. Russian propaganda on social media during the 2022 invasion of Ukraine. *EPJ Data Science*, 12(1): 35.

Golovchenko, Y. 2022. Fighting propaganda with censorship: A study of the Ukrainian ban on Russian social media. *The Journal of Politics*, 84(2): 639–654.

Guo, Z.; Zhang, Q.; An, X.; Zhang, Q.; Jøsang, A.; Kaplan, L. M.; Chen, F.; Jeong, D. H.; and Cho, J.-H. 2023. Uncertainty-aware reward-based deep reinforcement learning for intent analysis of social media information. *arXiv preprint arXiv:2302.10195*.

Habernal, I.; Pauli, P.; and Gurevych, I. 2018. Adapting serious game for fallacious argumentation to German: Pitfalls, insights, and best practices. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Haq, E.-U.; Tyson, G.; Lee, L.-H.; Braud, T.; and Hui, P. 2022. Twitter dataset for 2022 russo-ukrainian crisis. *arXiv preprint arXiv:2203.02955*.

Hernandez, J. 2022. Why Luhansk and Donetsk are key to understanding the latest escalation in Ukraine. https://www.npr.org/2022/02/22/1082345068/why-luhansk-and-donetsk-are-key-to-understanding-the-latest-escalation-in-ukrain.

Li, H.; Tomko, M.; Vasardani, M.; and Baldwin, T. 2022. MultiSpanQA: A dataset for multi-span question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1250–1260.

Miller, C. R. 1939. The techniques of propaganda. From "how to detect and analyze propaganda," an address given at town hall. *The Center for learning*.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Park, C. Y.; Mendelsohn, J.; Field, A.; and Tsvetkov, Y. 2022. Challenges and opportunities in information manipulation detection: An examination of wartime Russian media. *Findings of the Association for Computational Linguistics: EMNLP 2022*, 5209–5235.

Pohl, J. S.; Markmann, S.; Assenmacher, D.; and Grimme, C. 2023. Invasion@ Ukraine: Providing and Describing a Twitter Streaming Dataset That Captures the Outbreak of War Between Russia and Ukraine in 2022. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 1093–1101.

Röder, M.; Both, A.; and Hinneburg, A. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, 399–408.

Shevtsov, A.; Tzagkarakis, C.; Antonakaki, D.; Pratikakis, P.; and Ioannidis, S. 2022. Twitter Dataset on the Russo-Ukrainian War. arXiv:2204.08530.

Smart, B.; Watt, J.; Benedetti, S.; Mitchell, L.; and Roughan, M. 2022. # IStandWithPutin versus# IStandWithUkraine: the interaction of bots and humans in discussion of the Russia/Ukraine war. In *International Conference on Social Informatics*, 34–53. Springer.

Soares, F. B.; Gruzd, A.; and Mai, P. 2023. Falling for Russian Propaganda: Understanding the Factors that Contribute to Belief in Pro-Kremlin Disinformation on Social Media. *Social Media+ Society*, 9(4): 20563051231220330.

Torok, R. 2015. Symbiotic radicalisation strategies: Propaganda tools and neuro linguistic programming.

Zhou, X.; Shu, K.; Phoha, V. V.; Liu, H.; and Zafarani, R. 2022. "This is fake! shared it by mistake": Assessing the intent of fake news spreaders. In *Proceedings of the ACM Web Conference 2022*, 3685–3694.

AAAI ICWSM Paper Checklist

Written by the AAAI ICWSM 2024 Organizing Committee

pc.chairs@icwsm.org

Abstract

This document offers a Paper Checklist to be appended at the end of all submissions to, at a minimum, the September 2023 and January 2024 rounds of the AAAI ICWSM conference.

Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, our research contributes a valuable tool to the research community while upholding the principles of ethical and social responsibility. We have taken careful measures to ensure that our work does not violate privacy norms, perpetuate unfair profiling, exacerbate socio-economic divides, or show disrespect to any societies or cultures. For a detailed explanation of how we address these concerns, please refer to the Ethical Statement section of our paper.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, the Abstract and Introduction sections accurately reflect our paper's contributions and scope.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, in our paper, we have provided comprehensive explanations of our methodological approaches, ensuring they align with the claims we make.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? NA
   (e) Did you describe the limitations of your work? Yes, please see the Limitations section.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, in our research, we have thoroughly addressed potential negative societal impacts in both the Ethical Statement and Limitations sections. We acknowledge that sensitive and private concerns may arise, even with rigorous efforts to filter them out. To mitigate this, we have implemented a strict policy of conducting only aggregated statistical analyses. This approach minimizes the risk of revealing individual data or infringing on privacy, ensuring that our research is conducted with the utmost consideration for ethical and societal implications.
   (g) Did you discuss any potential misuse of your work? Yes, in the Limitations section of our work, we have addressed the potential for misuse. We emphasize that any future research leveraging our dataset should be mindful of the specific nature and context of the information collected.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, please see the Release and Access, Ethical Statement, and Limitations sections.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, we have read the ethics review guidelines and ensured that our paper conforms to them.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, we have ensured that the necessary resources for reproducing our experimental results are readily available. We have fine-tuned GPT models using OpenAI's official API, and have included comprehensive details of the queries and functions employed in our study in the Appendix.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, we have meticulously detailed all the training aspects of our study in the Weak Annotation section. This includes information about the data splits, the specific queries used, and the few-shot examples that guided our model training.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No, because our work primarily involved fine-tuning GPT models via OpenAI's official API. This process did not necessitate the use of external computing resources on our end.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, we address the potential risks and consequences of misclassification in our research in the Limitations section.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? NA

(b) Did you mention the license of the assets? NA

(c) Did you include any new assets in the supplemental material or as a URL? Yes, our paper includes URLs for accessing our newly released dataset. Additionally, we have provided the dataset as part of the supplemental material.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes, we have made our dataset available through Zenodo, a service that indexes datasets to enhance their findability. The link to our dataset is included in our paper for easy access. Additionally, we have detailed the access policy of our dataset. Please refer to the Release and Access and Ethical Statement sections of our paper.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? No, we did not create a separate Datasheet, because our paper comprehensively covers all the aspects recommended in the Datasheets for Datasets. This includes a thorough discussion on the motivations behind the data collection, the composition of the dataset, and an in-depth description of the data collection, processing, and annotation processes. Additionally, we have elaborated on the intended usage and distribution policies of our dataset, ensuring transparency and clarity in all aspects related to our dataset.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Yes, we have included the complete set of instructions given to annotators in the Human Annotation section of our paper. Additionally, we have provided a screenshot of the annotation platform in the Appendix.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, we confirm that all annotators were made aware of these risks and consented to participate. Furthermore, we have provided details of the Institutional Review Board (IRB) approval, ensuring ethical compliance, in the Ethical Statement section.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? No, we did not include the estimated hourly wage or total compensation for the annotators in our study. This is because the annotators were employed by the organization we collaborated with, and information regarding their wages is confidential and not available to our research team.

(d) Did you discuss how data is stored, shared, and deidentified? Yes, we have detailed in our paper how the dataset is shared in compliance with Twitter's Terms of Service, ensuring ethical data handling and distribution practices. Please refer to the Release and Access and Ethical Statement sections of our paper. In addition, we do not share any information that could potentially identify individual annotators.

References

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Appendix

Data Collection

Twitter Accounts Table 10 details the full list of government-affiliated Twitter accounts we use for data collection.

	Accounts
Russian Accounts	@RT_com, @redfishstream @tassagency_en, @Ruptly @SputnikInt, @RT_India_news @FiorellaIsabelM, @RTcreativeLab @RachBlevins, @stranahan @RuptlyVU, @Renegade_Inc @JaredDeLuna, @RT_webproducers @Romanovs100, @ICYMIvideo @russiabeyond, @BoomBustRT @RT_1917, @rt_play @RedactedTonight, @RTSportNews @WatchingHawks, @ilpetrenko_rt @MuradGazdiev, @ManilaChan @RuptlyUGC, @RT_PressOffice_ @WorldsApart_RT, @RT_Doc @PLCROSSTALK, @NastiaChurkina @MFinoshina_RT, @PaulaSlier @OksanaBoyko_RT, @seansparkthomas @RTUKnews, @RT_FreeVideo @QuestionMore, @RT_America @wyattreed13, @bamnecessary @GUnderground_TV, @afshinrattansi @TechUpdate_RT, @RTUKproducer @RSGovUK, @georgegalloway @IgorZhdanovRT, @Kosarev_RT @Sputnik_Insight, @DonaldCourter @FaultLinesRadio, @SophieCo_RT @zvezda_int, @NewsThing @RT_sputnik, @BreakingTheSet @NewswithEd, @InQuestionRT @SeanThomasRT, @SputnikMisfits @PortableTVApp, @LatviaSputnik @capitalfmmoscow, @AntonChesnokov1 @ruptly_newsroom
Ukrainian Accounts	@United24media, @MFA_Ukraine @DefenceU, @UKRinUN @UkrEmbLondon, @oleksiireznikov @SergiyKyslytsya, @EmineDzheppar @FedorovMykhailo, @OMarkarova @OlegNikolenko_, @VPrystaiko

Table 10: Full List of Russian and Ukrainian Government-Affiliated Twitter Accounts Utilized for Data Collection

Topic Modeling Tables 11 and 12 present the top 10 salient terms from each topic identified by the LDA topic modeling algorithm in tweets from Russian and Ukrainian accounts, respectively.

Topic Index	Topic Top-10 Salient Terms
1	ukraine, russia, say, kiev, conflict president, natio, military, ukrainian, zelensky
2	russian, putin, moscow, say, russia president, foreign, vladimir, minister, press
3	war, warn, crisis, year, energy world, power, ukrainian, moscow, people
4	ukrainian, russian, russia, report, case war, crisis, ukraine, center, day
5	ukraine, war, join, russia, watch london, live, biden, moat, uk

Table 11: Top 10 Salient Terms Identified in Each Topic Generated by LDA Modeling of Tweets from Russian Accounts

Topic Index	Topic Top-10 Salient Terms
1	ukraine, russian, support, invasion, armed russia, information, discuss, war, meet
2	russian, ukrainian, russia, standwithukraine ukraine, kill, region, city, missile, civilian
3	ukraine, russia, resolution, aggression, today unga, member, right, state, vote
4	ukrainian, ukraine, day, people, today world, russian, win, good, fight
5	thank, enemy, ukraine, slavaukraini, combat feb, loss, total, friend, ukrainian
6	ukraine, russia, crimea, day, war occupy, crimean, crimeaplatform russian, support
7	ukraine, ukrainian, russia, peace, security meeting, russian, war, support, live
8	ukraine, ukrainian, russia, people, diplomatic year, humanitarian, today, russian, support
9	ukraine, people, day, standwithukraine, wish kyiv, congratulation, russian independence, colleague

Table 12: Top 10 Salient Terms Identified in Each Topic Generated by LDA Modeling of Tweets from Ukrainian Accounts

Data Annotation

Pre-Annotation Table 13 outlines the queries utilized during the pre-annotation stage prior to human annotation. We specifically include ”provide exact text spans” to prompt the model to identify and highlight specific sections of the original text, rather than producing free-form responses.

	Pre-Annotation Queries to GPT-4
CTA	Is this tweet calling for any actions? Yes or no? *[If ”Yes” was provided, ask the following:]* 1. Who (provide exact text spans) are being called? 2. What actions (provide exact text spans) are being called?
DE	Are there any entities being discreditied in this tweet? Yes or no? *[If ”Yes” was provided, ask the following:]* 1. Who (provide exact text spans) are being called? 2. How (provide exact text spans) are they being discredited?

Table 13: Queries Used During Pre-Annotation

Machine Annotation Table 14 details the system message and queries used for the intent binary classification task in the machine annotation stage. These elements are uniformly applied across all three approaches we employ: zero-shot GPT-4-Turbo, GPT-4-Turbo with In-Context Learning (ICL), and Fine-Tuned (FT) GPT-3.5-Turbo.

	Queries
System Message	You are a social media content moderator that detect intent behind user posts.
CTA	Determin if the tweet contains a call to action. Format the response in JSON: {’label’: bool}. Tweet: {tweet}
DE	Determin if the tweet discredits any entities. Format the response in JSON: {’label’: bool}. Tweet: {tweet}

Table 14: Queries Used for Intent Binary Classification During Machine Annotation in Fine-Tuning, Zero-Shot, and In-Context Learning Approaches

	Queries
System Message	You are a social media content moderator that detect intent behind user posts.
CTA	The given tweet contains a call to actions. Pinpoint the called actions and the corresponding subjects. If no subject is explicitly named, leave the called_subjects field blank. Tweet: {tweet}
DE	The given tweet discredits some entities. Pinpoint each discredited entity and the corresponding phrases that discredit them. Tweet: {tweet}

Table 15: Queries Used for Span Localization During Machine Annotation in Fine-Tuning, Zero-Shot, And In-Context Learning Approaches

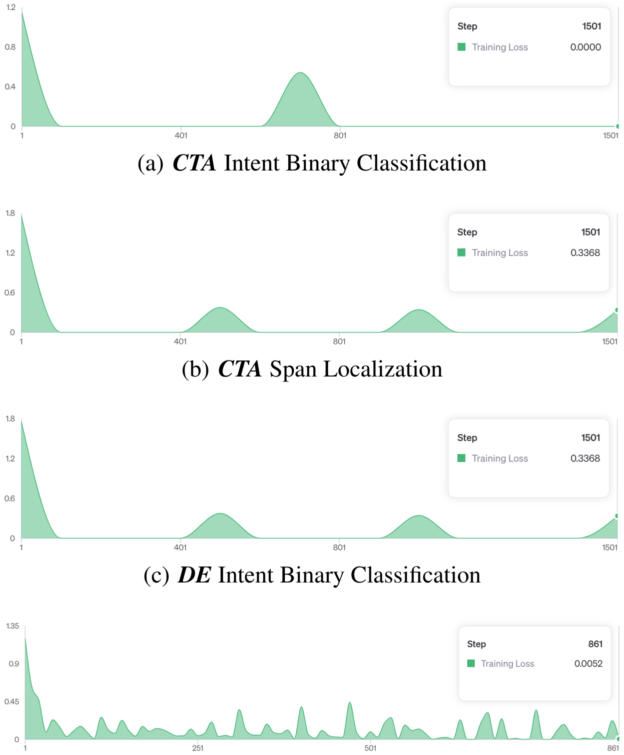
Table 15 details the system message and queries used for the interrogative span localization task in the machine annotation stage. These elements are uniformly applied across all three approaches we employ: zero-shot GPT-4-Turbo, GPT-4-Turbo with In-Context Learning (ICL), and Fine-Tuned (FT) GPT-3.5-Turbo.

The following code illustrates the JSON schema that we provide to the model for the function-calling process.

```
localize_call_to_action_function = {
  "name": "localize_call_to_action",
  "parameters": {
    "type": "object",
    "properties": {
      "called_subjects": {
        "type": "array",
        "items": {
          "type": "object",
          "properties": {
            "span_text": {"type": "string"},
          }
        }
      },
      "called_actions": {
        "type": "array",
        "items": {
          "type": "object",
          "properties": {
            "span_text": {"type": "string"},
          }
        }
      }
    }
  }
}

localize_discredited_entities_function = {
  "name": "localize_discredited_entities",
  "parameters": {
    "type": "object",
    "properties": {
      "discredited_entities": {
        "type": "array",
        "items": {
          "type": "object",
          "properties": {
            "span_text": {"type": "string"},
          }
        }
      },
      "discrediting_phrases": {
        "type": "array",
        "items": {
          "type": "object",
          "properties": {
            "span_text": {"type": "
                string"},
          }
        }
      }
    }
  }
}
```

Fine-Tuning GPT-3.5-Turbo We fine-tuned GPT-3.5-Turbo utilizing the OpenAI model fine-tuning API[3] for both the intent binary classification task and the interrogative span localization task. For both fine-tuning tasks, we prepared the training data in the same format as that used by OpenAI's Chat Completions API[4], employing the queries detailed in Tables 14 and 15 respectively. In both instances, we adopted a test/train split ratio of 3/7 and fine-tuned the model over 3 epochs. The training loss as a function of training steps across tasks is depicted in Figure 9.



(a) ***CTA*** Intent Binary Classification

(b) ***CTA*** Span Localization

(c) ***DE*** Intent Binary Classification

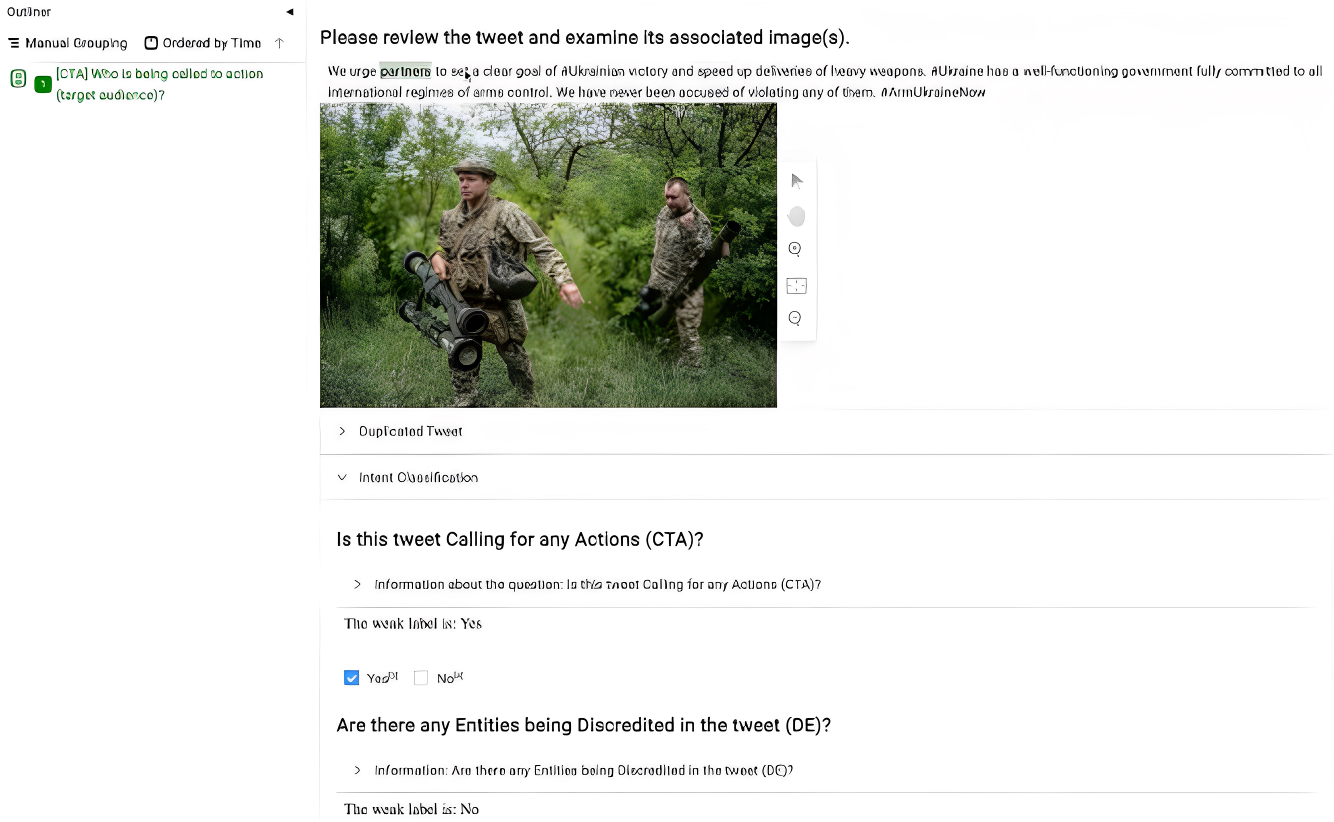
(d) ***DE*** Span Localization

Figure 9: Training Loss During the Fine-Tuning of GPT-3.5-Turbo Across Tasks

Human Annotation Platform Figure 10 displays the user interface of our custom annotation platform. The screenshot is intended solely for demonstrating the interface. To comply with Twitter's Developer Terms regarding redistribution and to prevent unintended disclosure, the text in the screenshot has been deliberately blurred and distorted.

[3] https://platform.openai.com/docs/guides/fine-tuning

[4] https://platform.openai.com/docs/api-reference/chat/create



Outliner

Manual Grouping

Ordered by Time

[CTA] Who is being called to action (target audience)?

Please review the tweet and examine its associated image(s).

We urge partners to set a clear goal of #Ukrainian victory and speed up deliveries of heavy weapons. #Ukraine has a well-functioning government fully committed to all international regimes of arms control. We have never been accused of violating any of them. #ArmUkraineNow

> Duplicated Tweet

⌄ Intent Classification

Is this tweet Calling for any Actions (CTA)?

> Information about the question: Is this tweet Calling for any Actions (CTA)?

The weak label is: Yes

☑ Yes[Y] ☐ No[N]

Are there any Entities being Discredited in the tweet (DE)?

> Information: Are there any Entities being Discredited in the tweet (DE)?

The weak label is: No

Figure 10: Screenshot of Our Annotation Platform: User Interface Demonstration with Text Intentionally Distorted