Identifying social media patterns in the Russia-Ukraine conflict NLP Course Project

Francesco Olivo
Master's Degree in Artificial Intelligence, University of Bologna
francesco.olivo2@studio.unibo.it

Abstract

This project examines the social media landscape during the early stages of the Russia-Ukraine conflict in March 2022. Utilizing a blend of Natural Language Processing (NLP), Machine Learning (ML), and social media analytics, the study aims to unravel the dynamics of social media interactions in response to sensitive geopolitical events. By analyzing patterns and behaviors manifested on various social platforms, the research endeavors to provide deeper insights into how global crises influence digital communication and public sentiment. The project's ultimate objective is to enhance our understanding of social media's role in reflecting and shaping public discourse during critical moments in international relations.

1 Introduction

In February 2022, the onset of the Russia-Ukraine conflict marked a significant moment in global geopolitics. This event not only intensified the public debate but also led to a notable polarization, reflecting a wide array of sentiments and ideologies associated with the two countries involved. Social media platforms, especially during the early stages of the conflict, became hotbeds for commentary, analysis, and the dissemination of diverse perspectives.

One notable aspect of this digital discourse was Russia's alleged efforts to influence Western public opinion through strategic social media interventions(NYT). This phenomenon raised intriguing questions about the detectability of such influences and patterns in social media narratives.

To explore these aspects, this project focuses on the analysis of social media content, specifically a dataset of tweets from the first week of March 2022. During this period, public attention and engagement with the Russia-Ukraine conflict were at their peak, providing a rich and significant dataset for analysis.

The methodology of this study involves two primary approaches. First, a text classification technique is employed, utilizing a fine-tuned model based on an existing text classifier(Li). This model is instrumental in identifying and categorizing the prevalent emotions expressed in each tweet, thereby offering insights into the emotional landscape of the social media discourse.

Second, to delve deeper into the structural aspects of social media interactions, a network analysis is conducted using NetworkX(Hagberg et al., 2008). This analysis focuses on the retweet network, aiming to uncover patterns of information dissemination and potential influences within the Twitter ecosystem during this critical period.

Through these methodologies, this project aims to shed light on the dynamics of social media behavior in response to global crises, specifically examining the interplay of emotions, information dissemination, and potential external influences in the context of the Russia-Ukraine conflict.

2 Background

In the realm of modern conflicts, social media has emerged as a pivotal battleground for information and influence. The ubiquity of platforms like Twitter and Facebook has transformed them into powerful tools for open-source intelligence (OSINT). Here, every tweet, post, or share becomes a potential data point in understanding the broader narrative of a conflict.

This phenomenon is not just limited to individual expressions. The aggregate data from social media provide a rich tapestry of public sentiment, making social media analytics a crucial aspect of conflict studies. Researchers and analysts use sophisticated tools to sift through vast amounts of data, identifying trends, sentiments, and potential misinformation campaigns.

The role of social media in conflicts extends beyond mere observation. It has become a tool for

state and non-state actors to influence public opinion, both domestically and internationally. In this context, understanding social media dynamics is not just about data analysis; it's also about discerning the intentions and strategies behind information dissemination.

This complexity is evident in the Russia-Ukraine conflict. The digital front of this conflict showcases a range of strategies, from straightforward propaganda to sophisticated disinformation campaigns. Analyzing these strategies through social media analytics provides invaluable insights into modern warfare's digital dimension.

Parallel to social media analytics, network analysis offers a unique lens to view the structural aspects of online interactions. At its core, network analysis in social media studies involves mapping and examining the connections between users and content. This method sheds light on how information, sentiment, and influence propagate through the social media landscape.

In the context of a conflict, network analysis can reveal the architecture of influence. By examining patterns of retweets, shares, and interactions, researchers can identify key nodes – users or accounts that wield significant influence over the spread of information. This analysis goes beyond the content of messages, delving into the mechanics of communication flow.

The utility of network analysis becomes particularly evident when examining coordinated influence campaigns. Such campaigns often leave distinct patterns in the network – clusters of accounts acting in concert, unusual propagation paths for certain narratives, and anomalies in user interactions. Identifying these patterns is crucial in understanding the mechanisms of digital influence in modern conflicts.

Together, social media analytics and network analysis form the backbone of understanding the digital landscape of contemporary geopolitical conflicts. They provide not just the 'what' of public sentiment and discourse but also the 'how' and 'why' of information spread and influence in the digital age.

3 System description

This project is an in-depth analysis of social media trends, specifically focusing on Twitter activity during the early stages of the Russia-Ukraine conflict. The methodology combines natural language processing and network analysis techniques, aimed at dissecting the complex web of public sentiment and information dissemination patterns that emerged on social media during this period.

The process begins with rigorous data preprocessing, a crucial step to ensure the integrity and clarity of the dataset. This dataset, comprising a significant volume of tweets gathered from the onset of the conflict, is subjected to a detailed cleaning process. This stage involves removing elements such as URLs, hashtags, user mentions, and stopwords, which are critical for refining the data. The aim is to strip away the noise and distill the essence of the textual content, ensuring that the subsequent analysis is based on relevant and uncluttered information.

At the heart of the project is the use of the DistilRoBERTa-base model for emotion classification. This model, known for its balance between computational efficiency and analytical accuracy, is adept at classifying text into various emotional categories. It has been trained to recognize Ekman's six basic emotions – anger, disgust, fear, joy, sadness, surprise – as well as a neutral state. This classification provides a foundational understanding of the emotional undertones prevalent in the social media discourse during the conflict.

A distinctive feature of our approach is the application of probability distributions for emotion analysis. Instead of assigning a single, definitive emotional label to each tweet, the model calculates the likelihood of each of the seven emotional states being present in the text. This method acknowledges the complexity of human emotions, particularly in a context as charged as a geopolitical conflict. It allows for a more nuanced understanding of the tweets, recognizing that a single message can convey a blend of emotions, each with varying degrees of intensity.

Building upon the emotional probabilities derived from the classification model, the project utilizes the K-means clustering algorithm to further segment the data. The decision to use seven clusters is a deliberate one, mirroring the number of emotions identified by the model. This clustering method groups tweets based on their emotional similarity, determined by the probability scores from the emotion analysis. This approach effectively segments the vast array of tweets into more manageable and thematically coherent subsets, facilitating a deeper and more focused analysis of

each emotional category.

Table 1 illustrates the composition of seven clusters based on the emotions detected in tweets. Cluster 0, dominated by Joy, suggests a positive or optimistic tone. Cluster 1, with high Disgust and Surprise, indicates reactions of shock or disapproval. Cluster 2, mostly Neutral, represents informational content. Cluster 3, marked by Fear, shows tweets expressing concern or apprehension. Cluster 4 is dominated by Anger, reflecting strong negative emotions. Cluster 5, characterized by Sadness, includes expressions of grief or sorrow. Finally, Cluster 6 shows a mix of Neutral and Joy, possibly indicating a balance of objective reporting and positive sentiment.

The final layer of our analysis is the network analysis, which examines the retweet patterns within the dataset. This part of the project goes beyond the content of the tweets and delves into the structural aspects of how information and emotions spread across the network. By mapping out and studying the interactions and influence patterns among Twitter users, the analysis identifies key influencers and nodes. These are the users or groups of users who play a significant role in shaping the discourse and propagating sentiments across the platform. This network analysis is crucial for understanding not just what is being said, but also how it permeates through the digital landscape of social media.

The project initially set out to analyze a substantial dataset of over 300,000 tweets from May 2023. The objective was to find a balance between the dataset's size and its relevance, especially considering the reduced volume of Twitter traffic on the topic in May 2023 compared to the initial months of the conflict. This period was selected with the expectation that it would provide a representative sample while managing the limitations of computational resources.

However, this initial approach encountered significant challenges. Primarily, restricting the dataset to English-language tweets from this specific timeframe had a pronounced impact on the study's results, especially in developing the social media network. One major limitation was the decision to rely exclusively on retweets, quotes, and replies as interaction metrics, overlooking "likes", which are often the most common form of engagement on Twitter. This approach resulted in a rather limited view of the interaction network, as evidenced by the largest connected component of the network comprising only 8 nodes. This size was insufficient for a comprehensive analysis of larger network dynamics, which are critical in understanding the full scope of social media interactions during such a complex geopolitical event.

To address these limitations, a revised approach was adopted. This subsequent method involved constructing the social media network first, focusing on identifying and analyzing tweets related to the significant nodes within this network. This network-centric approach allowed for a broader examination of the data, encompassing the entire month of May 2022. By focusing on tweets that were directly relevant to the network analysis, it became possible to include a wider time span and, consequently, a more diverse and potentially informative dataset.

Performing the emotion analysis after filtering tweets belonging to network nodes offered a more focused and relevant examination of the discourse. This method provided insights into how key players in the network shaped the conversation and the emotional tone of the discourse during the conflict. This approach also helped mitigate some of the computational challenges, as the analysis was concentrated on a more targeted subset of the data.

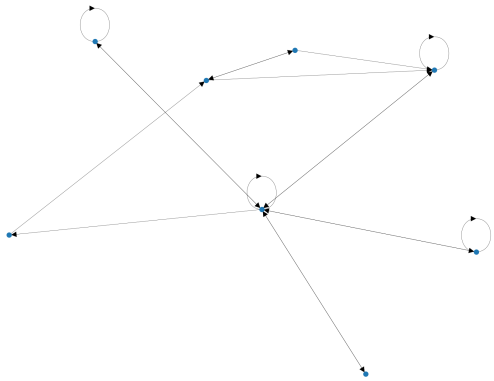

Figure 1: The largest connected network found for May 2023

Following the revised approach focusing on network analysis, the project successfully constructed a more expansive and informative social media network. This network, derived from the dataset encompassing the entire month of May 2022, comprises a substantial 4,089 nodes and 39,120 edges. This scale, as per empirical experience, strikes an optimal balance between providing significant in-

Cluster	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise
0	0.028	0.001	0.026	0.711	0.206	0.022	0.006
1	0.083	0.424	0.067	0.087	0.081	0.048	0.209
2	0.019	0.001	0.021	0.048	0.889	0.018	0.003
3	0.067	0.002	0.692	0.045	0.150	0.041	0.004
4	0.712	0.008	0.069	0.052	0.115	0.041	0.004
5	0.068	0.005	0.073	0.086	0.173	0.591	0.005
6	0.097	0.003	0.130	0.169	0.537	0.059	0.006

Table 1: Emotions clustering

sights and maintaining computational feasibility.

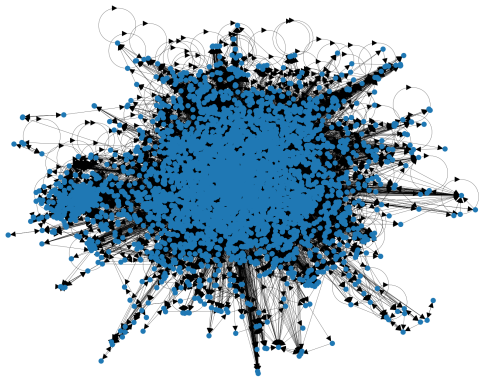

Figure 2: The new connected network found for May 2022

The network is designed as a directed graph, a structure that effectively captures the directional nature of interactions on Twitter. In this network, each user is represented as a node. Interactions, such as retweets and quotes, are represented as directed edges. These edges are drawn from the interacting user (the one who retweets or quotes) to the author of the original tweet. For instance, if user A retweets or quotes a tweet from user B, an edge is created that starts from node A and ends at node B. This directed edge effectively captures the flow of information or influence from one user to another.

This construction method allows for a nuanced understanding of the network dynamics. By mapping out these directed interactions, the network analysis can reveal not just who is talking about the conflict, but also how information and sentiments are being propagated through the Twitter ecosystem. It highlights who the key influencers are, who is being influenced, and how different users are interconnected in the discourse surrounding the conflict.

After successfully developing the social media network, the project progressed to a further level of analysis by clustering users within the network. This step was crucial for dissecting the complex interactions and identifying distinct patterns or "information bubbles" in the discourse.

The clustering of users was conducted based on two key metrics: the number of tweets each user posted and the distribution of sentiments in their tweets. This dual-criteria approach provided a multi-dimensional view of user behavior and sentiment expression within the network. The number of clusters was identified using the Elbow Method.

The primary goal of this clustering was to delve deeper into the network to uncover patterns and potential "information bubbles." Information bubbles, or echo chambers, are environments where users are primarily exposed to opinions and information that reinforce their own beliefs. Identifying these bubbles within the network is critical for understanding how different perspectives are formed and maintained on social media.

This clustering also aimed to discern if there were distinct groups within the network that consistently shared similar sentiments. Such groups might represent different factions or viewpoints in the discourse surrounding the conflict. By examining the clustering results, the project sought to gain insights into how information and emotions coalesce around certain nodes, potentially influencing the broader narrative.

4 Data

This project utilizes a comprehensive dataset consisting of tweets related to the ongoing Ukraine-Russia conflict(BwandoWando, 2022). The dataset is a result of systematic data collection from Twitter, providing an invaluable resource for analyzing

Cluster	N	Joy	Disgust_Surprise	Neutral	Fear	Anger	Sadness	Mixed_Emotions
1	0.00804	0.370	0.0603	0.357	0.0287	0.0162	0.0374	0.129
2	0.00277	0.0643	0.367	0.182	0.112	0.0253	0.148	0.101
3	0.0269	0.0932	0.0695	0.406	0.0445	0.0250	0.0440	0.318
4	0.0129	0.0397	0.0257	0.818	0.0155	0.00734	0.0204	0.0736
5	0.0277	0.113	0.0534	0.558	0.0286	0.0155	0.0399	0.191
6	0.0114	0.121	0.0293	0.157	0.0254	0.0126	0.0261	0.629

Table 2: Users cluster composition with respect to emotions

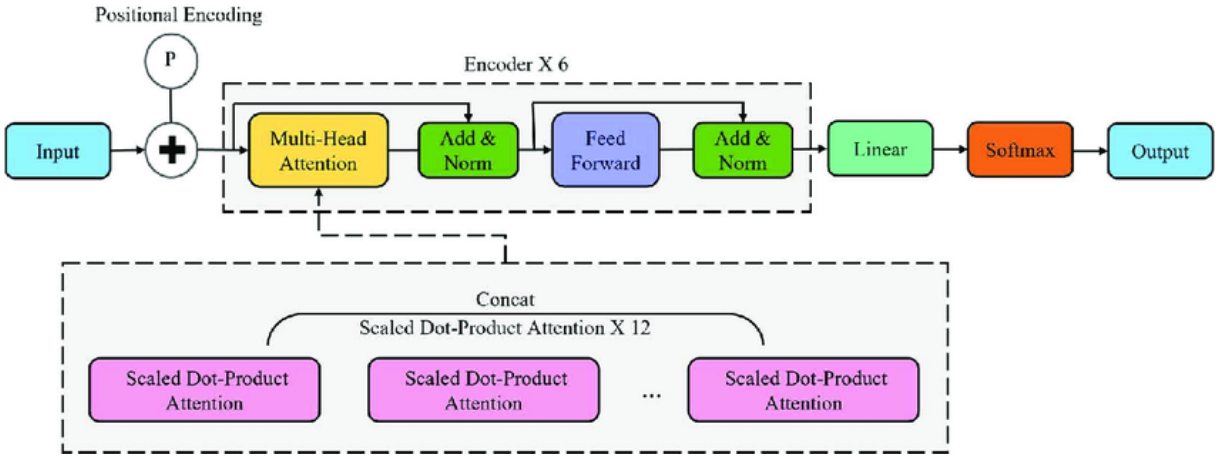

Figure 3: Model architecture(Gu et al., 2022)

public sentiment and communication patterns during the conflict. The dataset encompasses a wide array of tweets, including both original posts and retweets, collected over a significant period of the conflict. The data includes critical metadata such as timestamps and geolocation information, which are pivotal for temporal and spatial analysis. Additionally, fields like *is_retweet* and *original_tweet_id* are included to preserve the context of each tweet, allowing for a nuanced understanding of the discourse and the relationships between tweets. Data collection was performed using Python and the Tweepy library, with the support of services like Microsoft Azure and Anaconda Jupyter. The process involved dynamically querying Twitter's API to capture tweets related to the conflict, based on relevant hashtags, keywords, and geolocations. This approach ensured a comprehensive capture of the discourse surrounding the conflict, albeit within the constraints and limitations imposed by Twitter's API rate limits and terms of service.

5 Experimental setup and results

The experimental framework of this project centers around the fine-tuning of the DistilRoBERTa-base model, a process critical to adapting the model to the specific nuances of our dataset. The dataset in question includes a large collection of tweets pertaining to the Russia-Ukraine conflict, offering a rich source of natural language data for emotion classification.

For the fine-tuning process, we utilized the DistilRoBERTa-base architecture. The parameters for this phase were carefully selected to optimize the model's performance while balancing resource constraints. The fine-tuning was conducted over 10 epochs, with a specified 500 warmup steps to gradually ramp up the learning rate at the start of training. Additionally, we implemented a weight decay of 0.01 to help regularize the model and prevent overfitting.

The fine-tuning procedure was performed on a subsample of 10,000 instances drawn from our dataset. This sample size was chosen to strike a balance between achieving robust model performance and maintaining computational efficiency. For the validation process, a separate set of 1,000 instances was used to evaluate the model's generalization capability.

The training process yielded significant insights

into the model's performance. The final output of the training phase, as captured in the TrainOutput, showed a global step count of 6,250 and a training loss of 0.9867. These metrics indicate the extent of learning that the model underwent during the fine-tuning phase. Notably, the training runtime was approximately 2879.54 seconds, with the model processing about 34.73 samples per second and achieving a training steps per second rate of 2.17. The total floating-point operations (FLOPs) came to roughly 1.32e+16, a testament to the computational intensity of the training process.

Upon evaluating the model, the results painted a comprehensive picture of its performance. The evaluation yielded an accuracy of 21.06%, as indicated by the eval_accuracy metric. The loss during evaluation was 3.2604, higher than that in the training phase, which suggests challenges in model generalization to unseen data.

The results from the fine-tuning and evaluation phases suggest that while the model has learned from the training data, there are areas for improvement, especially in terms of accuracy and loss. The discrepancy between training loss and evaluation loss points to potential overfitting on the training data, a common challenge in machine learning models.

Phase	Loss	Accuracy	Runtime
Train	0.9867		2879.5426
Eval	3.2604	0.2106	100.4012

Table 3: Results of model training

6 Discussion

The users clustering allows for a deeper comprehension of the interactions within the network. The clusters described in table 2 can be summarised as following:

Cluster 1: The Balanced Emotional Range Balanced composition with significant proportions of joy, neutral, and mixed emotions, suggesting a diverse emotional response.

Cluster 2: The Reactive Discontent Characterized by high proportions of disgust and surprise, along with notable levels of sadness and fear, indicating more negative or reactive emotional content.

Cluster 3: The Diverse Emotional Spectrum Higher neutral and mixed emotions, with moderate levels of joy and disgust/surprise, suggesting a more varied or complex emotional response.

Cluster 4: The Factual or Unemotional Predominantly neutral, with very low levels of all other emotions, possibly representing more factual or unemotional content.

Cluster 5: The Positive Variety Dominance of neutral emotions, but also a reasonable presence of joy and mixed emotions, indicating a generally positive or varied emotional tone.

Cluster 6: The Emotional Ambiguity High proportion of mixed emotions, along with noticeable levels of joy, associated with tweets expressing ambiguity or a blend of different feelings.

With the successful development of the network and the clustering of users based on their tweeting behavior and sentiment expression, we have now progressed to a stage where it is possible to visually represent these clusters within the network. This visualization is a crucial step in our analysis, as it provides an intuitive and clear representation of the complex relationships and groupings within the Twitter discourse on the Russia-Ukraine conflict.

Visualization Description

The network visualization, as depicted in the figure below, illustrates the clusters formed based on the predefined criteria:

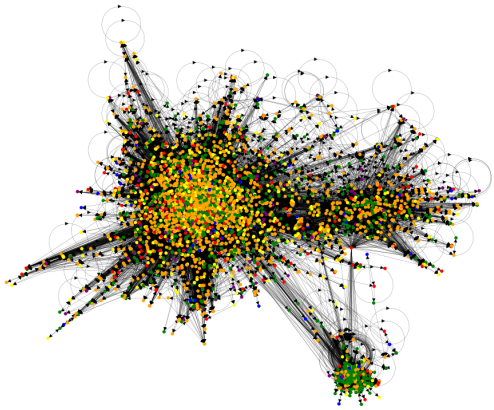

Figure 4: Clusters within the network

In this visualization, different colors represent different clusters, with each node in the network corresponding to a user and the edges representing interactions between them. A particularly no-

table observation is the pattern formed by the green nodes, representing Cluster 3.

Cluster 3, identifiable by its green nodes, exhibits characteristics of an echo chamber within the network. An echo chamber is a social structure where all members are exposed primarily to opinions that reinforce their own. The concentration and interconnectivity of green nodes suggest that users within this cluster are more likely to interact with and share content from users who have similar sentiment tendencies. This pattern is crucial for understanding the dynamics of information dissemination and sentiment reinforcement on social media platforms.

The analysis of the social media network includes the calculation of two critical metrics to understand the network's structure: the Homophily Index and the Assortativity Coefficient. These metrics are instrumental in gauging the extent to which users of similar types are connected.

A Homophily Index of 0.57 is observed in the network. This value indicates a moderate level of homophily, signifying that slightly more than half of the connections (edges) occur between nodes (users) of the same cluster.

Homophily in a network context refers to the propensity of individuals to associate with similar others. An index of 1 would represent perfect homophily, where all connections are between nodes of the same type, whereas an index of 0 suggests no homophily, with connections being completely random with respect to cluster. The index value of approximately 0.57 in this network denotes a balance: there is a tendency for similar users to connect, but it is not an overwhelmingly dominant feature of the network's structure.

The network's Assortativity Coefficient is measured at 0.25. This coefficient, ranging from -1 (perfect disassortativity) to 1 (perfect assortativity), evaluates the correlation of the cluster attribute across all connected nodes. A coefficient of 0.25 indicates a low to moderate level of assortativity, suggesting a tendency for nodes to connect with others of similar or the same cluster, although this tendency is not particularly strong. It represents a positive correlation but is not markedly high.

These metrics reveal the nuanced nature of the network. The moderate homophily index suggests a mix of diversity and like-minded interaction among users, potentially indicative of a variety of perspectives within the same cluster. The low to moderate assortativity coefficient points to a network that is not excessively compartmentalized by cluster, maintaining a degree of heterogeneity in connections.

Table 4 illustrates the echo chamber tendencies within various user clusters in the social media network, shedding light on how these clusters engage with content that aligns with their own viewpoints.

Cluster 5, labeled as "Positive Variety," shows the highest echo chamber tendency (0.4956), indicating a strong inclination among these users to engage with positively-toned content. Cluster 3, "Diverse Spectrum," with a moderate tendency score of 0.3856, suggests a mix of emotional content, albeit with a leaning towards similarity. In contrast, Clusters 2 and 4, named "Reactive Discontent" and "Factual/Unemotional" respectively, display low echo chamber scores (0.0728 and 0.0640), pointing to a broader range of content engagement, possibly encompassing diverse or neutral viewpoints.

The "Balanced Range" and "Emotional Ambiguity" clusters, with scores of 0.0450 and 0.0302, exhibit the least tendency for echo chamber behavior, suggesting these groups interact with a wide and varied range of sentiments and opinions.

7 Conclusion

The project's exploration of social media discourse during the Russia-Ukraine conflict has provided significant insights into user interactions and emotional expressions on Twitter, using advanced natural language processing and network analysis methods. This analysis has successfully delineated the emotional landscape and interaction dynamics in a substantial dataset of tweets, revealing diverse emotional expressions and interaction styles.

However, the study encountered specific challenges, including the limitations of focusing exclusively on English-language tweets and computational resource constraints. While these were partly addressed through adaptive strategies like a network-first analysis approach, they also highlight areas for potential improvement and further development.

A key area for future development could be the incorporation of multilingual data analysis to capture a broader spectrum of global perspectives on the conflict. Expanding the linguistic scope of the dataset would likely provide a more comprehensive understanding of the global discourse surrounding such events.

Cluster	Echo Chamber Tendency
5 (The Positive Variety)	0.4956
3 (The Diverse Emotional Spectrum)	0.3856
2 (The Reactive Discontent)	0.0728
4 (The Factual or Unemotional)	0.0640
1 (The Balanced Emotional Range)	0.0450
6 (The Emotional Ambiguity)	0.0302

Table 4: Echo Chamber Tendency by User Type

Additionally, the computational limitations encountered suggest the need for more robust data processing and analysis frameworks. Future iterations of this study could benefit from leveraging more powerful computational resources or optimizing existing algorithms for efficiency, enabling the analysis of larger datasets over extended time periods.

Furthermore, while the project successfully identified echo chamber tendencies within different user clusters, the approach could be refined to account for more nuanced aspects of user interactions, such as the impact of 'likes' and other forms of engagement beyond retweets and quotes. This would offer a more complete picture of user behavior and content dissemination patterns on social media.

In summary, this project has laid a solid foundation for understanding social media dynamics in geopolitical conflicts. Future work, building on this foundation and addressing the identified limitations and challenges, has the potential to offer even richer insights into social media's role in shaping public discourse and opinion in times of global crises.

8 Links to external resources

Git
Dataset

References

Inside russia's network of bots and trolls. https://www.nytimes.com/video/us/politics/100000005414346/how-russian-bots-and-trolls-invade-our-lives-and-elections.html.

BwandoWando. 2022. Ukraine conflict twitter dataset. https://www.kaggle.com/datasets/bwandowando/ukraine-russian-crisis-twitter-dataset-1-2-m-rows/data.

Yeonghyeon Gu, Xianghua Piao, Helin Yin, Dong Jin, Ri Zheng, and Seong Yoo. 2022. Domain-specific language model pre-training for korean tax law classification. *IEEE Access*, 10:1–1.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA.

Michelle Li. Fine-tuned distilroberta-base for emotion classification. https://huggingface.co/michellejieli/emotion_text_classifier.