Who did you choose?

A study on if there is a link between who we choose to support in different wars.

Case study: Russia-Ukraine and Israel- Palestine war (Reddit Version)

Babli Dey and Selenge Tulga

University of Rochester

Rochester, NY

{bdey2, stulga}@ur.rochester.edu

1 Abstract

This study investigates the dynamics of public support in the context of two contemporary geopolitical conflicts: the Russia-Ukraine war and the Israel-Palestine war. Utilizing Reddit as a platform for discourse analysis, we examine whether there is a discernible link between individuals' choices in supporting particular sides in these conflicts.

Our case study focuses on user comments related to these conflicts on Reddit. By leveraging natural language processing and stance analysis, we aim to classify expressions of support into categories such as Pro-Russia, Pro-Ukraine, Pro-Israel, Pro-Palestine, or Impartial.

In summary, our objective is to ascertain whether a correlation exists between individuals supporting a particular side in one conflict and their inclination to favor a specific side in a different conflict.

Keywords: Russia, Ukraine, Israel, Palestine, Reddit, comments, stance analysis, conflict, war, hama, gaza.

2 Introduction

In the internet age, social media platforms have become vital for global discourse, allowing people worldwide to express their views on various topics, including geopolitical conflicts, wars, human rights issues. And we see a surge of these opinions on social media since the Covid-19 pandemic. Reddit is one of those social media platforms where we saw the surge of opinions. Reddit, is a popular and influential social media platform where users participate in diverse discussions, share links, and post content spanning a wide array of subjects. With its diverse and engaged user base, Reddit serves as a central hub for fostering internet culture, facilitating discussions, and enabling the sharing of information and opinions on a multitude of topics[1].

Ever since the end of World War 2, there were many wars between many countries not like the scale of world wars but invading a country was a very pre 21st century thought. On 24 February 2022, Russia invaded Ukraine in an escalation of the Russia-Ukrainian War that started in 2014. The invasion was the biggest attack on a European country since World War II. The Russia-Ukraine conflict, characterized by territorial disputes, political turmoil, and military actions, has garnered significant international attention [2]. In contrast,

The Palestine-Israel conflict is a long-standing and deeply complex geopolitical struggle that has captivated global attention for decades. Rooted in historical, territorial, religious, and political disputes. Both conflicts have ignited passionate debates and drawn supporters and critics from across the globe.

3 Related work

As social media continues to gain prominence in people's lives, the information produced on these platforms has evolved into a valuable resource for researchers. "While there has been thorough investigation into two wars and reddit scraping [3], considerable attention has also been devoted to the analysis of stance [4]."

In the review of articles published between in 2021 [5], researchers examined 493,877 tweets using popular hashtags that studied specific to the conflict portraying opinions neutral or partial to Israel and Palestine. In this study, English-speaking countries exclude the valued opinions of non-English speaking countries, especially near Israel-Palestine which would have an impact on the study. Their study point out that there exists a class imbalance issue in the data, with pro-Israel tweets being very few in comparison to pro-Palestine tweets.

Our search aligns with the findings proposed in another study.[6], which uses reddit API PRAW to crawl the comments through the keyword-searching function in the package, so the comments collected must contain at least two keywords. As a result, they had 172,091 comments from 6,466 subreddits that were generated by 107,522 unique users, spanning from March 1, 2020, to December 15, 2020.

4 Methodology:

1. **Data Scraping:** We utilized the Reddit API to gather data on specific topics.
2. **Data Pre-Processing:**The raw dataset underwent preparation for analysis through preprocessing.
3. **Creating Comment Category:** Employing regular expressions, we categorized comments into wars based on word frequency derived from a Word Cloud.
4. **Stance Analysis:** Stance detection was conducted on a per-comment basis, involving two separate analyses for the Israel-Palestine and Russia-Ukraine conflicts..
5. **Inferencing the result:** A heat map was generated to discern any potential associations among war supporters based on the obtained results.

4.1 Data Scraping

For this study to collect comments, we utilized the PRAW Reddit API Wrapper, specifically the asynchronous version, async praw, a widely adopted Python package in Reddit-related research. To collect comments, we compiled a list of topics,which includes ['Russia,' 'Ukraine Conflict,' '.russia.ukraine.,' 'Ukraine,' 'Russian war,' 'Russia invades Ukraine,' 'Ukraine war,' 'UA,' 'RU,' 'Ukrainian,' 'Israel,' 'Palestine,' 'Israeli,' 'Israel,' 'Gaza,' 'Palestine,' and 'Hamas.']

To collect comments, we developed a function that searches for posts containing any of the words from the topics list. Using the post IDs of the matched posts, we then extracted comments that also included any of the keywords from the topics list. This process involved utilizing the subreddit and search functions within the PRAW package.

Field	**Description**
subreddit	Name of subreddit
commentId	Unique comment id
commentTimeStamp	Time stamp of comment
commentAuthor	Username of author
comment	Comment text
topic	Topic name

Table 1: Summary of the reddit fields

We collected 209,151 comments and 65,771 unique authors. Duplicate, nonEnglish, and automatic moderator comments are pruned from the dataset. Our final dataset contains 147,488 comments from 65,734 distinct authors.

Subreddit name	No.of comments	No. of Authors	Avg. Comment/Author
r/worldnews	50462	27309	1.8
r/europe	15560	6046	2.6
r/mapPorn	10622	4623	2.3
r/politics	7161	3590	2.0
r/news	5933	3240	1.8
r/CredibleDefence	4811	736	6.5
r/ukraine	2506	1900	1.3
r/facepalm	2418	1457	1.7
r/publicFreakout	2002	1181	1.7
r/coolguides	1290	1069	1.2
225	147,488	65,734	2.2

Table 2: The data collection

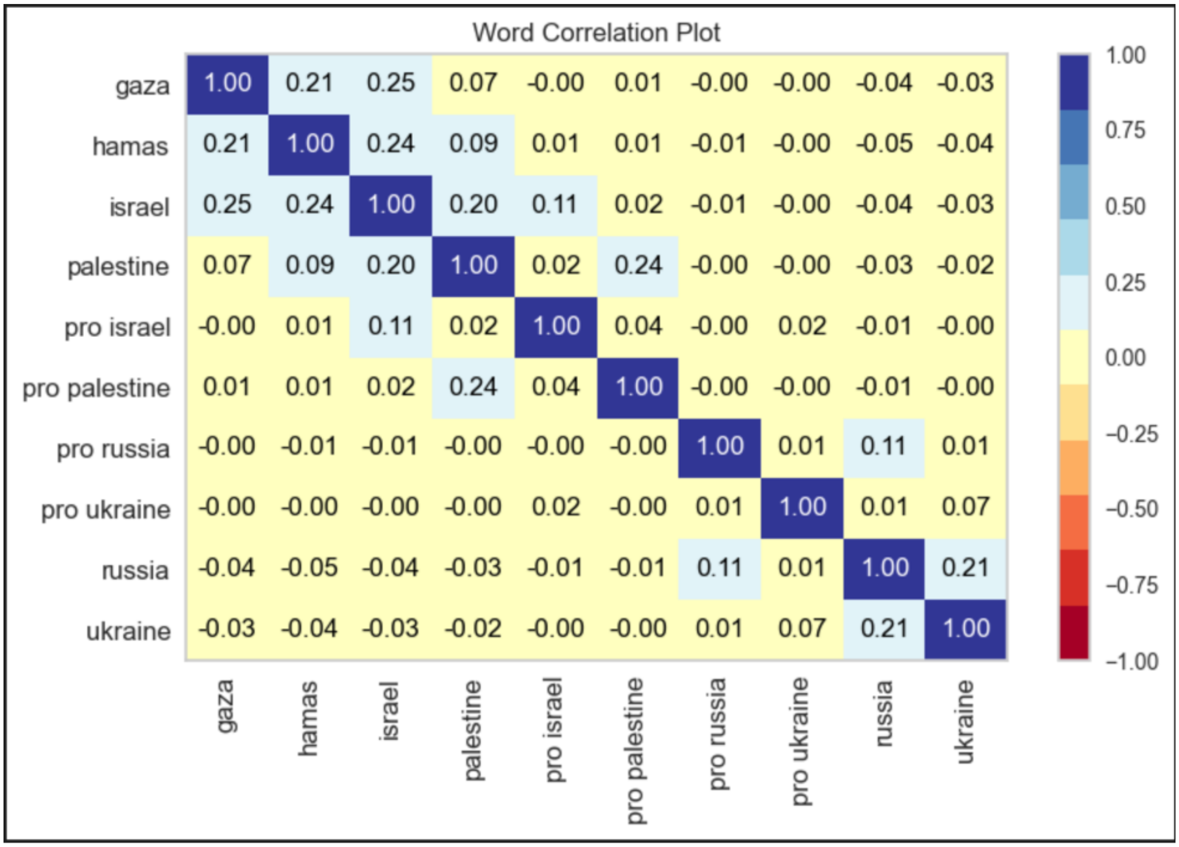


Figure 1: The word correlation plot

4.2 Data Preprocessing

To prepare the data for further stance classification and language modeling, we perform a text cleaning process. We convert all words to lowercase and remove all special characters, blank spaces and numbers from the text. Then we tokenize texts and remove stop words like “not”, “and”, “the”. Additionally we use lemmatization of words using NLTK and join the tokens back into a single string.

4.3 Creating Comment Category

Before establishing comment categories, we conducted word cloud analyses on both the original and preprocessed comments. This approach aimed to identify new word variations. Utilizing the word frequency insights, we developed a script employing regular expressions to categorize comments into four groups: Israel-Palestine, Russia-Ukraine, Both, and Other.

Comments falling under the Israel-Palestine category contained words related to either Israel or Palestine. Similarly, for the Russia-Ukraine category, comments featured words related to both conflicts. Comments falling into both categories were tagged as Both, while those not aligning with any of the conflicts were labeled as Other.

Given the substantial number of comments categorized as Other, we iteratively refined our regular expressions by manually inspecting and incorporating potential expressions into the script. This iterative process aimed to enhance the accuracy of categorization and reduce the number of comments labeled as Other.

4.4 Stance Analysis

To evaluate the stance of comments on both wars, we employed the Llama model [7] for training and prediction. The process involved multiple steps. Initially, we extracted a sample of 400 records from the dataset, utilizing random sampling across each comment category to create our training set. All the comments had to undergo a texting processing as the models had a character limit of 1024 characters. We utilized Excel formulas and Python scripts to process the text and obtain the desired format for modeling the training set.

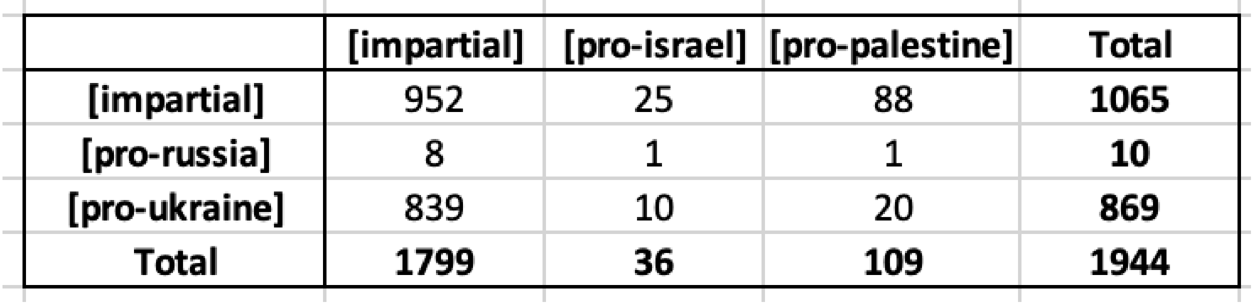
For the training phase, we formulated a prompt for ChatGPT 3.5, manually inputting comments for analysis. However, due to guideline and policy violations, not all comments yielded responses from ChatGPT. The prompt included two stance scores for each war: a score of -1 for pro-Russia, pro-Israel stances respectively, a score of 1 for pro-Palestine, pro-Ukraine stances respectively , and a score of 0 if the analysis indicated impartiality for both categories respectively.

After creating the training set, we had a sample dataset of 296 records with results from ChatGPT. To train our Llama-2-7B model, we utilized the Google Colab Pro subscription, to use Jupyter notebook due to its compatibility with the required configurations. The A100 GPU with 40 GB of GPU RAM was used for efficient processing. To make predictions, we excluded the sample training set from our main data set using comment IDs and randomly sampled 1000 comments from each comment category. The decision to limit the total comments to 4000 was influenced by time constraints associated with the execution time of the models.

4.5 Inference The Results

Among the 4000 comments, the Llama model successfully processed results for only 1944 comments, primarily due to the unavailability of the A100 GPU in Google Colab Pro. Consequently, our analysis is based on the 1944 comments that were successfully parsed.

From these 1944 comments, it was observed that 952 commenters expressed an impartial stance about both wars. Interestingly, the majority of comments tended to favor Ukraine, irrespective of the commenter's stance on the Israel-Palestine war. Similarly, in the context of the Israel-Palestine war, a significant number of comments favored Palestine, regardless of the commenter's stance on the Russia-Ukraine war.



	[impartial]	[pro-israel]	[pro-palestine]	Total
[impartial]	952	25	88	1065
[pro-russia]	8	1	1	10
[pro-ukraine]	839	10	20	869
Total	1799	36	109	1944

Table 2: results from the llama model (the numbers represent no. of comments)

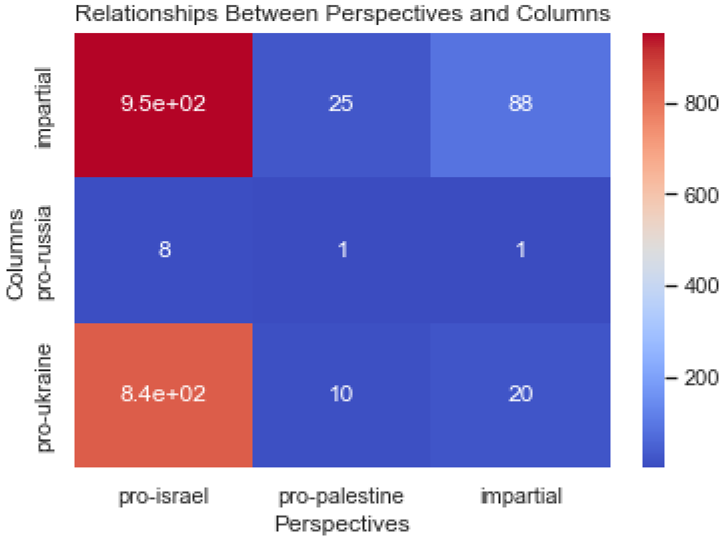


Figure 2. Heat Map of Result

Due to the limited data, we opted to conduct a Fisher's Exact Test to determine whether there is an association between comments supporting pro-Israel, pro-Palestine, pro-Russia, and pro-Ukraine stances

Ho: There is no association between the country preference (pro-russia/pro-ukraine) and the stance on the Israel-Palestine war.

Ha: There is an association between the country preference.

The values of the statistical inference: alpha = 0.05, p = 1 and odds ratio = 2

The calculated odds ratio of 2, coupled with a p-value of 1, suggests that there is no statistically significant association between individuals' country preferences (pro-Russia/pro-Ukraine) and their stances on the Israeli-Palestinian conflict (pro-Israel/pro-Palestine). The odds ratio indicates a twofold difference in supporting specific sides in both conflicts, but this observed difference lacks significance as the p-value surpasses the common threshold of 0.05. Consequently, the results do not provide compelling evidence to reject the null hypothesis, emphasizing the absence of a meaningful relationship in the given dataset.

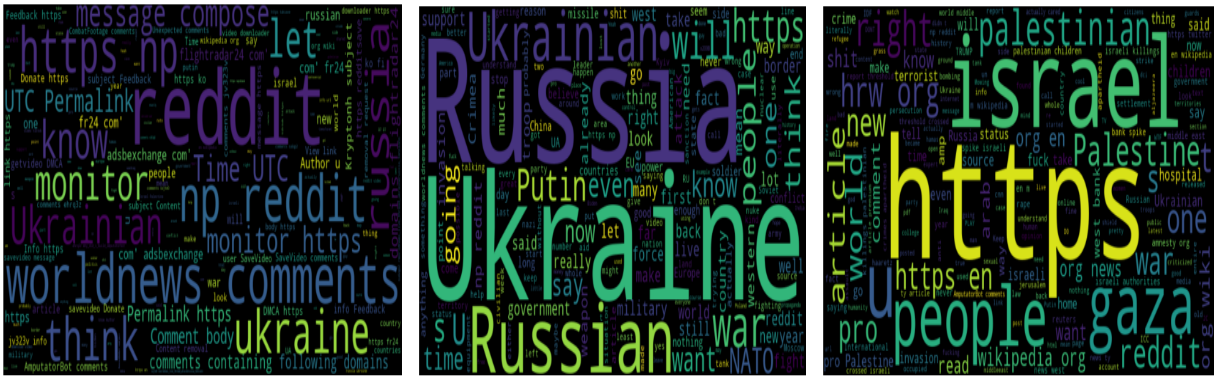


Fig 3. Word Cloud from Results- from left (impartial to both wars, only pro-israel, only pro-russia)

5 Experiment:

Sentiment Analysis

VADER [9] is a lexicon- and rule-based sentiment analysis tool that can handle words, abbreviations, slang, emoticons, and emojis commonly found in social media. It is typically much faster than machine learning algorithms, as it requires no training. Each body of text produces a vector of sentiment scores with negative, neutral, positive, and compound polarities. The negative, neutral, and positive polarities are normalized to be between 0 and 1. The compound polarity can be thought of as an aggregate measure of all the other sentiments, normalized to be between −1 (negative) and 1 (positive). In our study, As a result:
Negative 82,247 (55.7%), Positive 46,544 (31.5%), Neutral 18,696 (12.8%). In this case, over half of people have negative polarities for these conflicts.

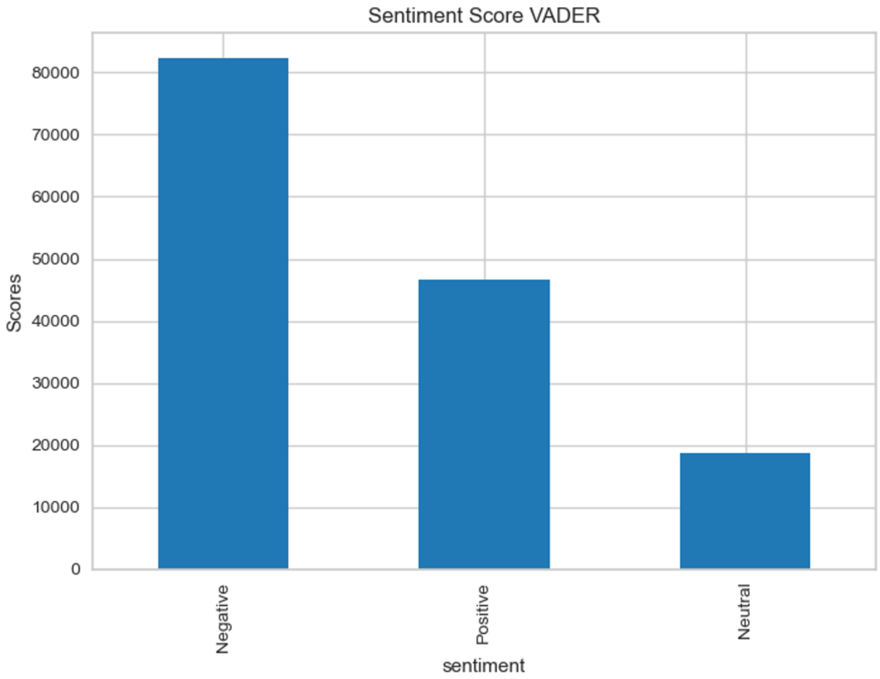


Figure 4: Sentiment Score graph

Topic Modeling

Common topics within the corpus were identified using latent Dirichlet allocation (LDA) [8], which is a common method used to identify themes in collections of scholarly textual data, as well as large sets of documents or texts in research or clinical contexts. LDA is a Bayesian probabilistic modeling method that aims to identify the unknown number of latent topics that are assumed to underlie a body of text. LDA draws from a Dirichlet distribution to generate distributions of probabilities that describe how words and documents are related to the latent topics underlying the dataset. Specifically, word-topic-probabilities are estimates of the probability a word is generated from a specific topic, whilst document-topic-probabilities are estimates of the probability that a topic has been generated in a specific document. Inspection of the highest word-topic and document-topic probabilities for each topic can help characterize the theme of each latent topic. In the current study, we used LDA to find common topics within the corpus of a user's comment from subreddits. Before feeding the comments into the LDA model, we remove punctuations from the com-ments and conduct lemmatization. To determine the most appropriate number of topics, we train LDA models with different numbers of topics using Gensim and graph their respective coherence scores. The highest coherence score (0.51) is reached when the number of topics is set to 5. However, after graphing the perplexity, we find optimal number 7. The coherence score of the chosen model is 0.46.

Topics	Keywords
Israel-Palestine conflict	jew, israel, people, palestine, jewish, arab, country, state, genocide, palestinian
Gaza Conflict and Civilian Casualties	hamas, israel, civilian, israeli, gaza, child, people, palestinian, terrorist, killed
Geopolitical Issues	crimea, donbas, india, ussr, sea, river, ru, oil, mandate, azov
Russia-Ukraine Conflict	ukraine, russia, russian, ukrainian, u, war, military, would, force, putin
Legal and International Relations	law, source, said, article, international, president, trump, official, un, news
General Opinions and Sentiments	people, dont, like, think, would, im, get, war, want, ukrainian
Comprehensive Israel-Palestine Discussion	israel, palestinian, hamas, gaza, palestine, israeli, would, state, west, land

Table 3: LDA topics

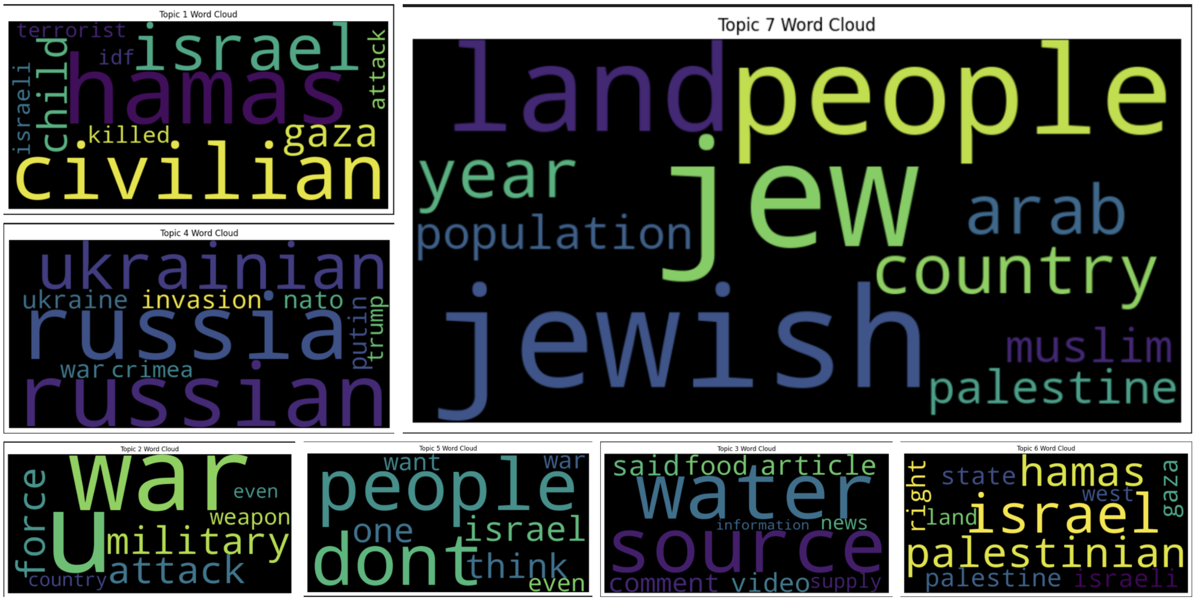


Figure 5: Topics generated from LDA

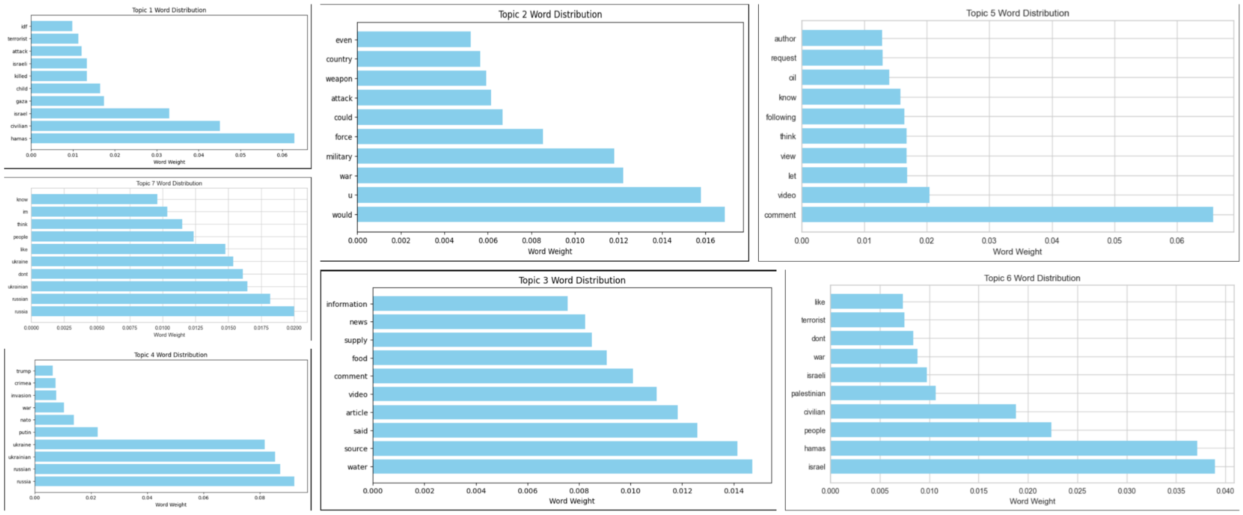


Figure 6: Word distribution generated from LDA

6 Conclusion

Despite the challenges faced, our analysis of the results suggests that there is no discernible connection between supporters of the two wars. Nevertheless, the llama model we constructed effectively assessed the stance of comments on each war. As a result, we recommend utilizing this model. It's important to note that our findings may be biased due to the encountered challenges, and a larger sample size, though attempted, was not attainable, potentially affecting the accuracy of our results.

7 Future work

In future work, employing the GPT model to gauge the stance in our analysis could be a valuable approach. Additionally, enhancing access to A100 GPUs or superior GPU capabilities would be beneficial for optimizing the functionality of the Llama model.

References

[1] Ahmed, Abubakar, Israel-Palestine Conflict: The World's Most Intractable Conflict (November 17, 2021).

[2] Liu Z, Shu M. The Russia–Ukraine conflict and the changing geopolitical landscape in the Middle East. China Int Strategy Rev. 2023

[3] Nathan Fox, Laura J. Graham, Felix Eigenbrod, James M. Bullock, Katherine E. Parks. Reddit: A novel data source for cultural ecosystem service studies. Ecosystem Services. Volume 50. 2021.

[4] D. KÜÇÜK and N. ARICI, "Sentiment, Stance, and Emotion Analysis on Twitter for COVID-19 Vaccination: A Survey", *AIS*, vol. 5, no. 1, pp. 14-22, Jun. 2022.

[5] Imtiaz, Arsal, Danish Khan, Hanjia Lyu and Jiebo Luo. "Taking sides: Public Opinion over the Israel-Palestine Conflict in 2021.

[6] Wu, Wei et al. "Characterizing Discourse about COVID-19 Vaccines: A Reddit Version of the Pandemic Story." Health data science vol. 2021 9837856. 27 Aug. 2021.

[7] Touvron, Hugo & Lavril, Thibaut & Izacard, Gautier & Martinet, Xavier & Lachaux, Marie-Anne & Lacroix, Timothée & Rozière, Baptiste & Goyal, Naman & Hambro, Eric & Azhar, Faisal & Rodriguez, Aurelien & Joulin, Armand & Grave, Edouard & Lample, Guillaume. LLaMA: Open and Efficient Foundation Language Models. 2023

[8] Westrupp EM, Greenwood CJ, Fuller-Tyszkiewicz M, Berkowitz TS, Hagg L, Youssef G. Text mining of Reddit posts: Using latent Dirichlet allocation to identify common parenting issues.2022

[9] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text", *ICWSM*, vol. 8, no. 1, pp. 216-225, May 2014.