Sentiment Analysis of Russo-Ukrainian War Using Social Media Discussion Corpus Applying NLP Techniques

Sumaiya Sinha
dept. of Computer Science Engineering
BRAC University
Dhaka, Bangladesh
sumaiya.sinha@g.bracu.ac.bd

Shouri Saha
dept. of Computer Science Engineering
BRAC University
Dhaka, Bangladesh
shouri.saha@g.bracu.ac.bd

Eftakhairul Islam
dept. of Computer Science Engineering
BRAC University
Dhaka, Bangladesh
eftakhairul.islam@g.bracu.ac.bd

Taslima Islam
dept. of Computer Science Engineering
BRAC University
Dhaka, Bangladesh
taslima.islam@g.bracu.ac.bd

Md. Sabbir Hossain
dept. of Computer Science Engineering
BRAC University
Dhaka, Bangladesh
ext.sabbir.hossain@bracu.ac.bd

Mehnaz Ara Fazal
dept. of Computer Science Engineering
BRAC University
Dhaka, Bangladesh
mehnaz.ara.fazal@g.bracu.ac.bd

Mr. Annajiat Alim Rasel
dept. of Computer Science Engineering
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—Sentiment analysis is an important technique in the Natural Language Processing field that makes computer software understand and analyze textual informationLATEX. Researchers use this method to better investigate and interpret human emotions and language to get valuable insights and information. It is growing more important as vast data is being spread through social media which can be made valuable for various industries or fields. Twitter and Reddit are two of the largest social platforms where valuable data is expressed by humans and can be easily accessed. So the aim of this study is to use this data to better understand human sentiments regarding conflicts. So that we can make NLP and sentiment analysis effective in these situations to better interpret conflicts. In this research we analyze Russian Ukrainian war through sentiment analysis by using the data available on Twitter and Reddit. Here we show that using various NLP models like ROBERTA, BOW, Decision tree we can interpret the positive and negative and neutral sentiment of the public. This study summarizes multifaceted discourse on this war and provides a comprehensive understanding.**

***Index Terms*—Sentiment Analysis, Russo-Ukrainian War, Twitter, Reddit, Social Media Data, Conflict, ROBERTA, Bag-of-Words (BOW), Decision Tree, Public Sentiments, Geopolitical Conflicts, Textual Information, Data Mining**

I. Introduction

Online social networks, also known as OSNs, were developed with the purpose of connecting people in order to facilitate the sharing of their feelings, desires, achievements, hobbies, and interests as an essential component of their socialisation. [7] Emotions are an inherent aspect of human nature. Consequently, they enhance our comprehension of our experiences and enable us to respond correctly. Comprehending emotions is crucial for comprehending human behavior. The proliferation of digital technology in the modern era has significantly altered the manner in which individuals communicate their thoughts and opinions on global current affairs. Social media platforms play a crucial role in facilitating the sharing of thoughts. Platforms such as Twitter, Reddit, and Facebook serve as prominent sources of information and potent tools for molding public opinion. Sentiment analysis has emerged as a significant tool in comprehending the emotions and viewpoints of social media users in this particular environment. Being cognizant of your emotions facilitates the process of determining your needs and desires. The reason for this is that possessing emotional awareness enables us to articulate our emotions with greater efficacy, mitigate or resolve conflicts, and navigate challenging emotions more effortlessly.[6]

This paper is a comprehensive analysis of the 2022 Russo-Ukrainian confrontation. The Russo-Ukrainian War has sparked a global debate and generated a vast amount of social media discussion. During the conflict, people from various regions experienced a range of emotions and held a variety of perspectives regarding the event. Through posting their thoughts and feelings on various social media platforms, individuals were able to express and share their opinions and feelings. Inspiration was drawn from a recent scholarly article concerning sentiment analysis in Covide-19. Our primary objective is to analyse and describe the sentiment associated with the intensification of the conflict on a particular online social network: Twitter. We are examining the tweets that consist just of text and have been posted by individuals globally during a designated time frame. Through the utilisation of the Bow model in addition to the Robustly Optimised BERT prior training Method (RoBERTa) model. We will classify tweets and posts under this dataset into three parallels: Negative, Positive and Neutral based on the sentiment scores assigned by the model. By conducting this study and writing the report, we will be able to provide evidence of what people all around the world think about the conflict, as well as how the war is contributing to the spread of negativity among people and how people are suffering as a result of it. So, this analysis will assist us in gauging public opinion on the subject and gathering data on the suffering of the populace, thereby contributing to the effort to bring attention to the matter. This research paper will contribute to the ever-growing field of sentiment analysis and has dwelled deep into the process of how the dataset was generated and how the classification models classify a highly engaged topic consisting of a plethora of sentiments.

To what degree, given the limits of each approach, can RoBERTa and BOW models effectively capture the growing sentiment towards the Russo-Ukrainian War on Twitter, and how can this knowledge be used to address real-world difficulties and guide policy decisions?

II. Literature Review

The following section summarizes Twitter and Reddit sentiment analysis studies on the Russia-Ukraine crisis. This review will inform our research design and analysis by highlighting earlier research's limits, methods, and findings.

A paper, ”Sentiment Analysis On Twitter Posts About The Russia and Ukraine War With Long Short-Term Memory” by Allwin M. Simarmata, Anthony, Tiffany, Matthew Evan Phanie (1) uses Dataset: 2537 Indonesian-language Twitter posts pertaining to the Russo-Ukrainian War were gathered. Preprocessed to eliminate extraneous data, stemmed, and tokenized. Weighted numerical vectors via TF-IDF conversion. Results: 54.7% expressed a positive sentiment, 35% a neutral sentiment, and 10.2% a negative sentiment.

A paper, ”“Sentiment of the Tweets on Russo-Ukrainian War: the Social Network Analysis” by A. Poleksić and S. Martinčić-Ipšć (2) uses datasets that include ”Ukraine” and ”Russia” tweets from May, October, November, and December 2022. The tweets were mostly negative (57%), neutral (32%), and positive (11%). Network research showed that #StopTheWar: Promoting peace and support for Ukraine and #StandWithUkraine focused on specific aspects of the conflict. Putin and #Russia: Supporting Russia. #Sanctions and Economy Economic effects of the war. Community attitude varied by ”Stanciness”: #StopTheWar and #StandWithUkraine was positive. Many positive and negative comments about Putin and Russia. Sanctions and Economy: Economic concerns have fuelled unfavourable sentiment.

A paper, “Public Sentiment and Emotion Analyses of Twitter Data on the 2022 Russian Invasion of Ukraine” by M. B. Garcia and A. Cunanan-Yabut (3) uses dataset: 27,894 tweets published on the initial day of the invasion that included the hashtag #UkraineRussia. The results were: 70% of the messages expressed a negative sentiment, 26% a neutral one, and 4% a positive one. The sentiments that were most prevalent were sadness (35%), anxiety (22%), anger (18%), and surprise (14%). The terms ”peace,” ”war,” ”people,” and ”the world” exhibited the highest frequency of occurrence among the keywords. The most frequently used hashtags included #StopTheWar, #StandWithUkraine, #NoWar, and #RussiaUkraineWar.

A paper, “Characterizing the 2022- Russo-Ukrainian Conflict Through the Lenses of Aspect-Based Sentiment Analysis: Dataset, Methodology, and Key Findings” by M. Caprolu, A. Sadighian and R. Di Pietro (4) uses a dataset: 5.5+ million conflict-related Twitter posts from 1.8+ million unique users. Results: Public opinion changed: Shock and support for Ukraine were followed by desensitization and conflict duration and impact concerns. Different components had different sentiment: More favorable about Ukrainian resistance and Western backing, negative about Russian aggressiveness and humanitarian crisis.

A paper, “Semantic Analysis of Russo-Ukrainian War Tweet Networks” by Benjamin Džubur, Žiga Trojer and Urša Zrimšek Russo-Ukrainian War (5) uses datasets of hashtag tweets that exceed 1.5 million tweets. The network found that prominent communities discussed #StandWithUkraine and support the Ukrainian people. #StopTheWar: Peaceful conflict resolution. Countering Russian propaganda. Economic sanctions against Russia. Finding community themes with LDA analysis increased public discourse. #StandWithUkraine wants peace, fears Ukraine, and condemns Russia. #StopTheWar: Demands urgent ceasefire, global concerns, and humanitarian relief. Fake news and manipulation are exposed by #RussianPropaganda. An assessment of sanctions, their economic effects, and calls for more. Community and issue trends were identified using network visualization.

This study adds to social media data sentiment analysis and global crisis emotion recognition literature. This study investigates popular sentiment and emotional reactions to events like the 2022 Russian invasion of Ukraine to better understand their social and psychological effects. Conflict resolution and global peacebuilding data may inform research, policy, and communication.

III. Methodology

A. Data collection

The Ukraine-Russia conflict has been ongoing since 2014 and it sparked many social media wars since then. To capture the sentiment on this conflict we picked two major social media platforms to collect data from. A vast amount of people went on Twitter and Reddit to express their thoughts and sentiments during this time. So we targeted Twitter and Reddit to get real-time updated data, community discussion, hashtag hashtag-focused posts. Because of global reach, we were able to bring variation in data as well as in language.

First, we picked a dataset from Kaggle that sources from Twitter which mainly captures everyday updated data regarding Ukraine-Russia. It gets updated every day but at the time of collection for this study, it contained 229813 Twitter posts. This vast dataset was collected in CSV format and then converted to a compressed file.

Another dataset ”Ukraine Russia War Reddit Data” was collected from Kaggle to access the data from Reddit as well. This dataset was made using a web scraping API named ”PRAW API”. This contains the post title, content, date, and author. This dataset was made in CSV format and contains 16708 posts regarding the Ukraine-Russia conflict.

Both of the datasets contain expressed sentiment in text format with other information. We further analyze and visualize data so that it can provide a clear understanding of the datasets.

B. Analysis of data

In the twitter dataset we analyze the csv files and get the shape. We will use only English language data. It contains 98422 columns and 17 rows out of 28 rows. It contains columns userid', 'username', 'acct desc', 'location', 'following', 'followers', 'Total tweets', 'usercreatedts', 'tweet id', 'tweet createdts', 'Retweet count', 'text', 'hashtags', 'language', 'coordinates', 'favorite count', 'extractedts'. As, we observe that the tweets are from various languages. So to train the model properly we focus on English and drop the others. We also further analyze the hashtags to extract information, and we see that here average num characters are 204.3 and average num words are 28.689210074984047 and median num characters are 211.0 and median num words are 29.0.

C. Data cleaning

Upon completion of data collection, the primary task is to do data cleansing. In this step, the process entails choosing the data that is pertinent to the given task. During the process of data cleaning, we have eliminated duplicate and irrelevant data, rectified structural mistakes, discarded extraneous features that are unrelated to our task, eliminated all emojis, addressed missing data, and ultimately verified for null values. Thoroughly filtering the data is crucial to guarantee that the model is trained on pertinent and precise information.

D. Pre-processing of text dataset

Data Pre-processing: Prior to training the model, the data must undergo pre-processing. This stage entails data cleansing, eliminating any extraneous information, and converting the data into a format that is appropriate for model training. Initially, we have extracted the English language data from the collection and exclusively focused on it, disregarding all other language datasets. The technique involves converting all the texts to lowercase, eliminating any URLs, deleting HTML tags, stripping punctuation from the messages, and neglecting retweets.

E. Classification Analysis

1) Sentiment Analysis (by using RoBERTa): We were able to examine the sentiment and emotions of the entire world by importing all English data from posts, comments, and captions on major social media platforms (such Reddit, Twitter, and so on).

2) Emotion Analysis (by using RoBERTa): We were able to examine the general emotions that various individuals in the dataset displayed.

F. Algorithm Description

1) RoBERTa in sentiment analysis:

1) As part of data preprocessing, online material like tweets and articles about the conflict are gathered. Cleaning, tokenization (word splitting), and lemmatization (base form reduction) are all preprocessing steps.
2) Embedding with RoBERTa: The RoBERTa model is told to handle each token by making a vector representation that captures its meaning and context. These vectors hold information about the word itself, how it connects with nearby words, and how it makes you feel in general.
3) Classification of mood: The encoded vectors are fed into a classifier that can figure out the mood, like a linear layer. The classifier guesses how the text feels and usually puts it into one of three groups: neutral, positive, or negative.

2) BOW in sentiment analysis:

1) Data Preprocessing: Text is lemmatized, sanitized, and tokenized, similar to RoBERTa. Stop words (such ”the” and ”a”) are removed.
2) Each word in the preprocessed text is converted into a BoW ”feature” during feature building. Frequency counts determine each word's feature value.
3) Sentiment Classification: A sentiment classifier receives a vector of word frequencies, the BoW representation. The classifier predicts positive, negative, or neutral text sentiment.

G. Twitter Hashtag Analysis and Displaying Wordcloud

Hashtag Analysis and Worcloud displayal are a crucial part of this research and will be discussed with graphical representation in the results section for a better understanding of the experiments conducted in our research to give an output. The wordcloud depicts those words that were mostly used in tweets all over the world, with the most used word as bigger fonts and least used word with smaller font. Whereas, the hashtag analysis is a bar chart that shows the word count of different hashtags used in tweets.

H. Result visualization

To conduct a proper sentiment analysis we are planning to use two dynamic and highly accurate algorithms, namely RoBERTa and BOW. RoBERTa excels at sentiment analysis of public opinion and emotional reactions to the Russia-Ukraine war. On the other hand, BoW directly and successfully analyzes Russia-Ukraine conflict public opinion. It delivers excellent insights about the conflict's digital discourse's language and sentiment patterns despite its simplicity. However, as of now we have not yielded any results yet and are planning further to get an accurate output in the near future for our next draft.

I. Ethical Considerations

For this research, we used publicly available social media posts and scrapped them from twitter and avoided acesseing data illegally. In our dataset we identified each user with an id and avoided showing names to keep anonymity. And we used the collected dataset just for analyzing purposes and maintained confidentiality and integrity of the data But since the posts are from the general public the data could be biased depending on the collected data.

J. FAIR and CARE Principles

our research aligns with the FAIR and CARE principal in the following ways:

1) Dataset collection were from diversified tweets (different geographic location, different demographic, millions of users). Also during data collection, we tried to minimize as much bias as possible.
2) During Model development, we used two recognized algorithms specifically made for sentiment analysis, namely, RoBERTa and BOW. We also fine-tuned our model to give a better output with a wide range of kaggle dataset. 80 % data was for training and 20 % data was for testing.
3) We tried to present our results and findings through concise and clear graphical representations to avoid any confusion.
4) Finally, we have kept guidelines for ethical considerations of sensitive information during the implication of our experiment.

IV. Result

A. Figures and Tables

TABLE I
Result Table

Result Name	**Comparison**			
	Algorithm	***positive***	***negative***	***neutral***
Sentiment	***RoBERTa***	***3rd***	***2nd***	***1st***

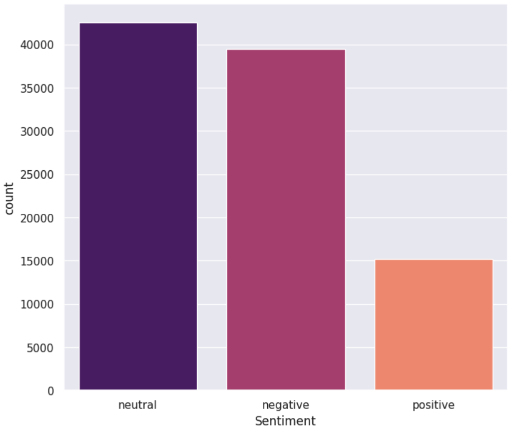

Figure Labels [Figure 1]: By inputting all english data from various social media (twitter, reddit, etc.) posts, comments, and captions, we were able to analyze the emotion and sentiment of the whole world. As you can see, most of the tweets had a negative sentiment amongst them regarding the Ukraine-Russia war though highest is neutral but the negative were close to highest. This shows that people were empathizing more with Ukraine and wanted Putin to stop the war. From the total dataset, around 39000 data were of negative sentiment which is close to maximum, 42000 were of neutral sentiment and a very few of them (14000) had positive sentiment within them, showing that some people supported the war.

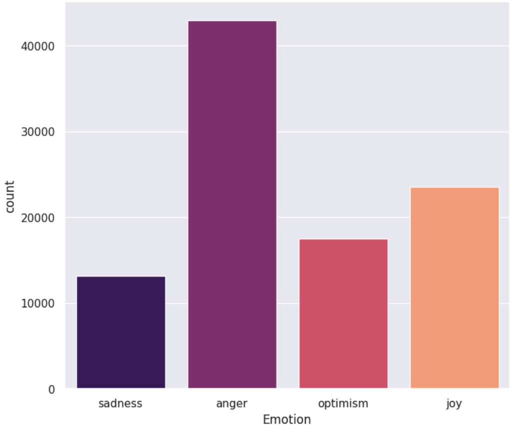

Figure Labels [Figure 2]: From this graph, we were able to analyze the overall feelings exhibited by different people throughout the datasets. As we can see, the maximum posts (44000) showed anger in them. This showed that most of the people were angry because of this war caused by Russia. Some data around 18000 expressed optimism, which could mean people were hoping that the war would be soon over. Furthermore, some people (13000) showed sadness throughout this post showing how saddening the consequences of this war was and a very few people, around 22000 were showing joy due to this war.

Limitations

Our research has some limitations that can improved in the future works. Our dataset contains multiple languages but in our research we dealt with English tweets only as specified models always doesn't work best on all languages. Our dataset may seem biased depending of the collection location. We used BOW and ROBERTA for analyzing the conflict sentiment. But more efficient model could be made building on this.

Future Works

1) Expand the analysis beyond English to include other prominent languages discussed online, such as Russian, Ukrainian, and other European languages.
2) Analyze the evolution of sentiment over time, identifying trends and shifts in public opinion as the conflict unfolds.
3) Identify the key topics discussed in social media conversations about the war, including specific events, actors, and narratives.
4) Compare and contrast the sentiment expressed towards the war across different social media platforms and geographical regions.
5) Develop explainable AI models for sentiment analysis to understand the factors that contribute to the sentiment expressed online.

Conclusion

Using Twitter and Reddit datasets, this study examined the sentiment of online discussions concerning the Russia-Ukraine conflict of 2022. Through the utilization of sentiment analysis methodologies such as RoBERTa and BoW, we acquired significant knowledge pertaining to the sentiments and opinions of the general public regarding the conflict.

The primary conclusions drawn from the analysis indicate that anxiety and anger were the prevailing emotions in English tweets, whereas a strong negative sentiment was observed in Russian tweets. Support for Ukraine, appeals for peace, and apprehension regarding the repercussions of the conflict emerged as prevalent global themes. The examination of texts in their original languages unveiled subtleties in both tone and lexical application.

References

1) Guerra, A., & Karakuş, O. (2023). Sentiment analysis for measuring hope and fear from Reddit posts during the 2022 Russo-Ukrainian conflict. Frontiers in Artificial Intelligence, 6. https://doi.org/10.3389/frai.2023.1163577
2) Xu, A., Tiffany, T., Phanie, M. E., & Simarmata, A. M. (2023). Sentiment analysis on Twitter posts about the Russia and Ukraine war with long Short-Term memory. Sinkron Jurnal Dan Penelitian Teknik Informatika, 8(2), 789–797. https://doi.org/10.33395/sinkron.v8i2.12235
3) A. Poleksić and S. Martinčić-Ipšć, ”Sentiment of the Tweets on Russo-Ukrainian War: the Social Network Analysis,” 2023 46th MIPRO ICT and Electronics Convention (MIPRO), Opatija, Croatia, 2023, pp. 1089-1095, doi: 10.23919/MIPRO57284.2023.10159770.
4) M. B. Garcia and A. Cunanan-Yabut, ”Public Sentiment and Emotion Analyses of Twitter Data on the 2022 Russian Invasion of Ukraine,” 2022 9th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), Semarang, Indonesia, 2022, pp. 242-247, doi: 10.1109/ICITACEE55701.2022.9924136.
5) M. Caprolu, A. Sadighian and R. Di Pietro, ”Characterizing the 2022- Russo-Ukrainian Conflict Through the Lenses of Aspect-Based Sentiment Analysis: Dataset, Methodology, and Key Findings,” 2023 32nd International Conference on Computer Communications and Networks (ICCCN), Honolulu, HI, USA, 2023, pp. 1-10, doi: 10.1109/ICCCN58024.2023.10230192.
6) “Understanding Your Emotions (for Teens) - Nemours KidsHealth.” https://kidshealth.org/en/teens/understandemotions.html (accessed Nov. 30, 2022).
7) Caprolu, M., Sadighian, A., & Di Pietro, R. (2023, July). 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) (pp. 1-10). IEEE.