National College of Ireland

Opinion Mining using Twitter data for Ukraine-Russia War

MSc Research Project
Data Analytics

Daksh Sharma
Student ID: 22165665

School of Computing
National College of Ireland

Supervisor: Mrs. Harshani Nagahamulla

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: .....Daksh Sharma .....

Student ID: .....22165665.....

Programme: .....Data Analytics..... **Year:** ...2023-2024

Module: .....MSc. Research Project.....

Supervisor: .....14/12/2023.....
Submission Due Date: .....

Project Title: ..... Opinion Mining using Twitter data for Ukraine-Russia War .....

Word Count: ...7167..... **Page Count**.....22.....

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....Daksh Sharma.....

Date: .....14/12/2023.....

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Opinion Mining Using Twitter Data for Ukraine-Russia War

Daksh Sharma
x22165665@student.ncirl.ie

Abstract

The conflict between Ukraine and Russia (RUW) escalated in 2022 but has been a topic of interest since 2014. Social media platforms like Twitter contain raw data that can be utilised to gain knowledge about public opinion and attitude towards a particular topic. In this study, tweets relating to the Ukraine-Russia conflict were obtained and used to provide insights about public opinion in the year 2023. The results were compared to the results obtained in previous studies until 2022 to dig deeper into the public opinion and figure out whether the sentiments of the people have shifted or remained constant. The main objective of this study is to effectively utilize BERT model to perform opinion mining on tweets relating to RUW. Our approach involves using VADER and RoBERTa to label the tweets and then use the labelled data to fine-tune a BERT model. In the second approach, the results from both VADER and RoBERTa were merged to create a dataset with true sentiment labels, which was used to fine-tune another BERT model. RoBERTa based BERT was found to be performing better than VADER based BERT with an average accuracy difference of 33.44%. This gives us the idea that transformers-based models are more effective in performing sentiment analysis than rule-based approaches.

Key word – Opinion Mining, Sentiment Analysis, BERT model, VADER, RoBERTa model, Ukraine-Russia War.

1 Introduction

The development of real-time information networking platforms such as X (formerly known as Twitter), has resulted in an unprecedented repository of public opinions on numerous worldwide entities that influence and effect human lifestyles. While X is an effective tool for expressing and creating ideas, it also offers new and distinct obstacles. Addressing these issues necessitates the use of powerful algorithms that can swiftly evaluate and grasp the vast array of ideas published on the platform (Bello et al. 2023a). Tweets, in contrast to more formal linguistic styles, have an informal linguistic style, with misspelled words, casual grammar usage, URL links, user mentions, hashtags, and other components. These basic characteristics create both obstacles and opportunities for machine learning and natural language processing (NLP) activities such as Sentiment Analysis (Barreto et al. 2023a). Sentiment analysis technologies have shown to be the most effective method of determining individual attitudes and feelings over time.

Sentiment analysis is a discipline that uses technological resources to investigate, interpret, and identify hidden emotions, feelings, and sentiments in text or interactions. It uses machine

learning (ML), natural language processing (NLP), data mining, and artificial intelligence (AI) methodologies to mine, extract, and categorize user's opinions on entities such as companies, products, individuals, services, events, or ideas, identifying various sentiments. Often referred to as opinion or sentiment mining, sentiment analysis captures the polarity of text, categorizing it as positive, negative, or neutral. There are primarily four types of sentiment analysis namely, Fine-grained, Aspect-based, Emotion detection, and Intent analysis [1]. Given the data has three sentiment classes namely, positive, negative, and neutral, fine-grained sentiment analysis is performed in which the aim is to categorize the tweets in one of the aforementioned sentiment categories.

The need for computers to understand human spoken and written language drives the demand for natural language processing (NLP). As a result, the bag-of-words (BoW) technique was developed which employs N-grams. However, BoW models neglect the contextual meaning of words. However, this model is dependent on a large corpus and significant computing power. Word2Vec was then created, which produces a single vector embedding for every word. Its limitation is that it only considers the left or right context. Google resolved these issues and enhanced language processing in 2018 with the introduction of a transformer model. In addition to tackling transfer learning problems successfully, it has shown notable success in a number of natural language processing tasks, such as sentiment analysis, question answering, and named entity recognition (Bello et al. 2023a).

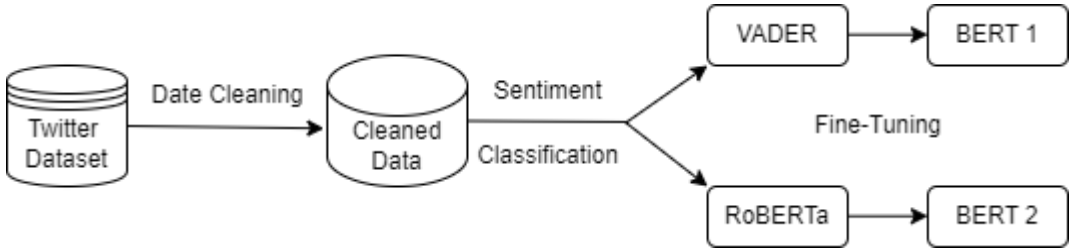

Figure 1 – Project Flowchart

The research question for this project is – **"How effectively can BERT models be applied to perform opinion mining on a Twitter dataset related to the Ukraine-Russia war with respect to accuracy?"**.

This study focuses mainly with exploring the efficient use of BERT models for opinion mining on a Twitter dataset relating to the Ukraine-Russia war. The inherent restrictions posed by the dynamic and informal structure of Twitter data gave birth to this research question. My goal is to find and improve the precision and comprehensiveness of sentiment analysis on tweets about this geopolitical event by utilizing BERT's advanced capabilities.

Figure 1 represents the steps undertaken in the project. In the upcoming sections, previous studies relating to this project has been described. The Related Work section aids the project by highlighting key points from previous studies, figure out gaps in the studies, and give an opportunity to innovate and bring novelty to this project. The methodology section describes the data-processing steps and the model building process. The Evaluation sections describes the key results of our study, and the Conclusion section encapsulates the results and future work for this project.

[1] https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-sentiment-analysis/

2 Related Work

The conflict between Russia and Ukraine has caused major disruptions to the global economy, escalated geopolitical tensions, and resulted in a substantial humanitarian crisis due to the influx of refugees. The conflict's far-reaching effects outside the surrounding region are highlighted by the effects it has on energy markets, security issues, and international relations. The study conducted by (Ahmed et al.) examined how the European stock market reacted to Russia's recognition of Donetsk and Luhansk, highlighting differences by country and industry. (Sohag et al.) examined food inflation in both Eastern and Western Europe, establishing connections between it and geopolitical threats, Russia's actions, and global worldwide energy prices. They also suggested policy measures for improved resilience. An additional investigation performed by (Sasmoko et al.) in their study, evaluated the environmental impact of the conflict and finds link between global carbon emissions and military operations between Russia and Ukraine, confirming the claim of ammunition emissions. Since the war between Ukraine and Russia has massive geopolitical and socio-economic impact globally, the analysis of these studies encouraged me to look into this topic and find out how the public is affected by this crisis and what are their views on this ongoing event.

The best way to get insights about public opinions is Sentiment Analysis. It is important because it draws conclusions from text and measures opinions, feelings, and attitudes of the general public towards a particular event (in our case – Ukraine-Russia war). This helps analysts, researchers, and corporations to understand public sentiment and make well informed decisions. There are several ways to perform sentiment analysis, including Machine Learning models, Natural language processing, Deep learning models, Lexicon-based approaches, and Hybrid approaches. Hybrid approach and novel models which were fine-tuned on specific dataset or are combinations of different deep learning models gave outstanding results. The study conducted by (Aslan et al.), proposed a unique Multistage Feature Extraction using Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) – MF-CNN-BiLSTM model which combined various qualities and benefits of CNN and BiLSTM. Another unique approach was proposed by Vyas et al., in which they utilized the combination of Emotion Robustly Optimized Bidirectional Encoder Representations from the Transformers Pre-training Approach (Emoroberta) and machine-learning techniques. (Wadhwani et al.) compared the accuracies of multiple machine learning models including extra trees classifier, random forest, logistic regression, support vector machine and many more with extra trees classifier (ETC) performing the best. In a separate study, (Barreto et al.) noticed that Emo2Vec, w2v-Edin, and RoBERTa were the best performing models with different sets of combinations of classifiers such as SVM and MLP. Nandurkar et al. had a similar approach in comparing multiple Naïve Bayes models such as MultinomialNB, BernouliNB, and GaussianNB.

BERT is a pre-trained natural language processing (NLP) model created by Google. It can handle long-range dependencies effectively as it is based on the Transformer architecture and uses a bidirectional method that captures information from both sides of a word. It was pre-trained by unsupervised learning and anticipates missing words in phrases to acquire contextualised word representations. BERT can be fine-tuned for downstream tasks such as sentiment analysis. Li, in his study, fine-tuned the pre-trained BERT model for the sentiment analysis task, but found better results with RoBERTa which suggests that a BERT model with an improved pre-training approach is better than fine-tuning a normally pre-trained

BERT model for downstream tasks. (Bello et al.) proposed a combination of BERT with CNN, RNN, and BiSLTM. The proposed BERT was used to generate a vector according to the word context and the deep learning classifiers were used to predict the sentiment output. In the study conducted by (Sirisha and Chandana), they proposed a unique and hybrid approach to perform aspect-based sentiment analysis by combining RoBERTa and LSTM, utilising their advantages and fixing their respective flaws. La Gatta et al. in their study, utilised the capabilities of twitter data and variations of BERT model on a separate task of retrieving false claims on Twitter. They chose RoBERTa transformer model to extract the vector embeddings of both claims and tweets because of the higher similarity scored in comparison to other transformer models like ms-marco-MiniLM-L-4-v2 and quora-roberta-base. The limitation of BERT model to process a maximum of 510 tokens at a time, encouraged (Sheng and Yuan) to create BERT-based fusion model utilising BiGRU network to solve the problem of long text sentiment classification on Chinese data. Their approach combines the BiGRU network to comprehend complicated structures in Chinese articles, uses BERT-based models as base classifiers to capture partial sentiments, and aggregates output from N BERT-based models. Results of these studies implies that utilising the capabilities of BERT model is one of the best approaches for our opinion mining study.

In the study conducted by (Ramos and Chang), DistilRoBERTa variant and pre-trained XLM-RoBERTa-Base model were used to perform emotion classification. English tweets were categorised into seven different emotion categories while Russian tweets were classified into positive, negative, and neutral sentiment. In separate research by (Thakkar et al.), they utilised the capabilities of BERT model and its variations on a larger scale by working on a dataset of 1.5 million tweets in English and 13 other languages. They compared the results of fine-tuned BERT on English tweets with M-BERT model, which was trained on a corpus of 13 different languages supported by Text Blob. We know that utilising multilingual data in the model improves the versatility of the model and promotes greater generalization. It also ensures inclusion by avoiding biases towards specific languages and cultures, resulting in more equitable and accurate sentiment prediction. (Dominic et al.) in his study presents another way to handle multilingual data by translating the tweets to English using neural machine translation. A special model to perform sentiment analysis on Bangla language YouTube comments data was introduced in the study by (Hasan et al.) called BanglaBERT. Hyperparameter optimization was performed on five models including XLM-RoBERTa and Distilm-BERT. On comparison, BanglaBERT outperformed all the other transformer-based classifiers.

Since twitter was rebranded as X in 2023, twitter does not allow users with twitter developer account to extract tweets using Twitter API with a basic account anymore. You need an account with a monthly subscription to extract tweets using API but with number of tweets restricted to 5000 per month. Most of the studies conducted on similar topics used the publicly available datasets or extracted tweets using web scraping tools from the year 2022. This adds to the novelty of our work since we are taking tweets only from the year 2023. This will help us to compare the results of similar studies from 2022 and determine if the sentiments and feelings of the public has shifted in the following year. With the state-of-the-art results provided by BERT and its variations, it is clearly the best model choice for our study.

3 Methodology

3.1 Data Description

The data set used for this study was the Russia-Ukraine War (RUW) tweets which is available for download on Kaggle (https://www.kaggle.com/datasets/towhidultonmoy/russia-vs-ukraine-tweets-datasetdaily-updated/). It contains the tweets collected on 27th and 28th February 2023 related to Ukraine-Russia War. The author collected a total of 10,014 tweets. The data set contains a total of 36 columns including, date, time, username, name, tweet, and language. Since the main focus of the study is sentiment analysis and keeping data regulations in mind, only the 'tweet' column is filtered out.

3.2 Data Pre-processing

In order to have a strong and effective analysis framework, it is essential that data be cleaned and pre-processed. By reducing data dimensionality and noise, these crucial steps play an important role in increasing the efficacy and speed of machine learning and deep learning models. The complex nature of web text data, which is primarily comprised of unstructured forms, creates a number of issues. This data frequently contains informal language and a variety of symbols, as well as segments of data that are either uninformative or irrelevant to the research. Tackling these difficulties through rigorous cleaning and pre-processing paves the path for a more streamlined and successful analytical procedure. This helps to improve the quality of input models and creates an environment that is favourable for deriving valuable insights from the large quantity of data obtained from web text data (Dominic et al. 2023).
In the study conducted by (Thakkar et al. 2023), the data pre-processing steps for sentiment analysis task involves removing emojis, converting to lowercase, eliminating links, mentions, non-UTF8/ASCII characters, and disregarding hashtags for enhanced text consistency.

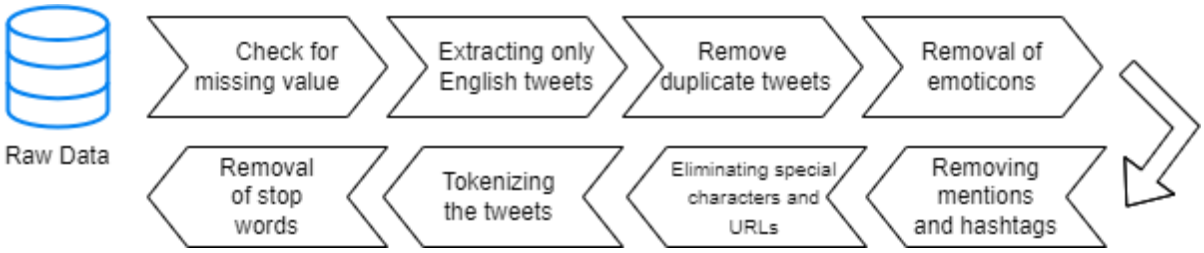

Figure 2 – Data Pre-processing steps

The data set was checked for missing values during the first stage of data preprocessing. Although a number of variables, including 'location', 'quote_url', and 'source', had null values, our main focus column 'tweet', showed none. As a result, no special handling of missing values was needed. A new data-frame was created that only included tweets in English language. 49 duplicate entries were identified and subsequently removed. Using a predetermined Unicode pattern, emoticons were identified and then deleted from the dataset. All tweets were converted to lowercase, and the 'emoticons' column was removed, in an effort to improve text uniformity.

Next, terms that started with '@', which usually indicated mentions from other users on Twitter, were removed. Subsequent improvements involved eliminating hashtags, special characters, and URLs using a function on the 'tweet' column was tokenized using the word tokenize function from the NLTK (Natural Language Tool Kit) package to make further analysis easier. Additionally, a collection of English stop-words extracted from the NLTK dataset was utilized to eliminate frequently occurring, uninformative terms. Ultimately, these stop words were eliminated from the tokenized text using the function "remove_stop_words", which greatly improved the dataset for sentiment analysis on tweets pertaining to the conflict between Russia and Ukraine (Refer to figure 2).

3.3 Sentiment Analysis

3.3.1 VADER

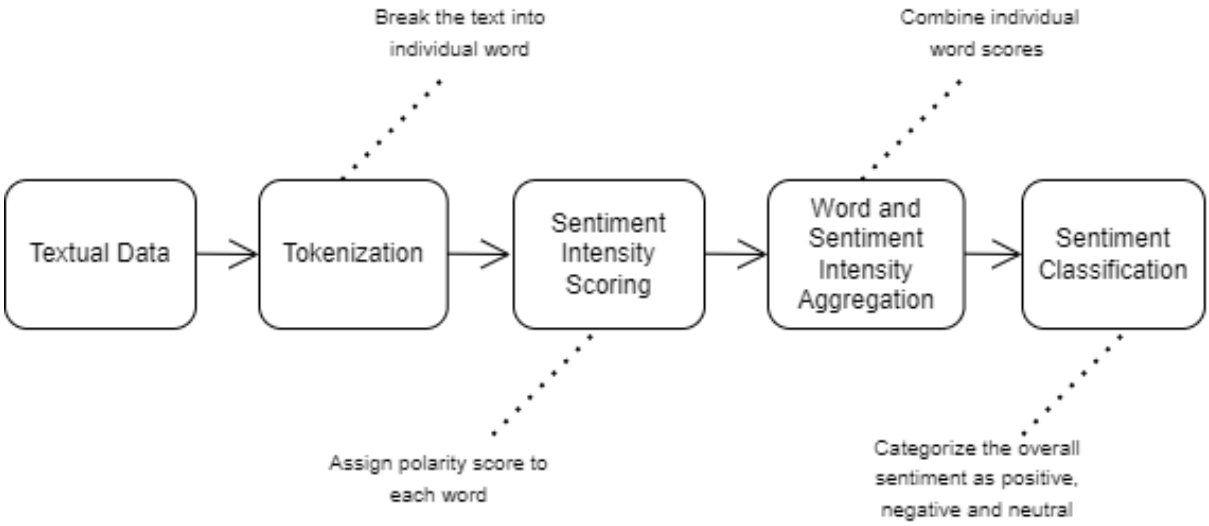

Figure 3 – VADER sentiment analysis

VADER stands for Valence Aware Dictionary and sEntiment Reasoner. This sentiment analysis tool was created specifically for social media text, especially brief and informal content like tweets. VADER is a sentiment analysis tool that has been trained on a large corpus of words and phrases with corresponding sentiment scores by researchers at the Georgia Institute of Technology (Hutto and Gilbert 2014). In order to operate, VADER analyses the text for sentiment-expressing words and phrases. It then assigns polarity scores to those words and phrases based on the language's context and intensity. In order to better understand the complexities of sentiment in informal text, it considers elements such as negations, capitalization, and punctuation. VADER can be applied to the tweets data in context of our research on the conflict between Ukraine and Russia to get sentiment scores belonging to one of the three sentiment categories. The compound score, which is a normalized weighted composite score, can be used to categorize tweets as positive, negative, or neutral.

First, an instance of the SentimentIntensityAnalyzer from the vaderSentiment library named 'analyzer' is created which is designed for sentiment analysis in textual data. It assigns a compound sentiment score to input text using pre-built sentiment lexicons, representing the

overall sentiment. In order to obtain the sentiment scores, we iterate through each row of the data frame, extracting tweets, and then apply the VADER's polarity_scores method. The sentiment analysis results are added to the dataset, and the final data frame has extra columns for compound, positive, negative, and neutral sentiment scores. We use the sentiment scores ('Compound') from the VADER Sentiment analysis and define criteria for classifying tweets as 'Negative', 'Positive', or 'Neutral'. The resulting 'Category' column is a simple representation of tweet sentiment, with a default 'Neutral' label for unclassified cases. This improvement makes it easier to explore sentiment classes inside the data, making the analysis more comprehensible and insightful.

3.3.2 RoBERTa

RoBERTa stands for Robustly Optimised BERT Approach. RoBERTa is a pre-trained language model developed by Facebook AI. It is based on the transformer architecture and is an optimised variation of BERT model, with a few new adjustments [2]. While BERT has two pre-training objectives namely, Masked Language Model (MLM) and Next Sentence Prediction (NSP), RoBERTa simplifies the pre-training objective by only using the MLM task. RoBERTa is trained on a larger corpus of data than BERT (Liu et al. 2019). The model used in this part of the project is Twitter-roBERTa-base for Sentiment Analysis provided by Hugging Face. This model was trained on 58 million tweets and fine-tuned for sentiment analysis with TweetEval benchmark [3]. The output of this model is a sentiment score dictionary for each tweet. Each tweet gets a set of values of predicted probabilities for it to be positive, negative, and neutral. Higher value of positive score, negative score and neutral score indicates higher predicted probability of a positive, negative, and neutral sentiment respectively, for a particular tweet. To use this output for sentiment analysis, we extracted the sentiment label based on the highest probability. For example, if 'roberta_pos' has the highest value, then the tweet is classified as Positive, and similarly for 'roberta_neg' and 'roberta_neu'.

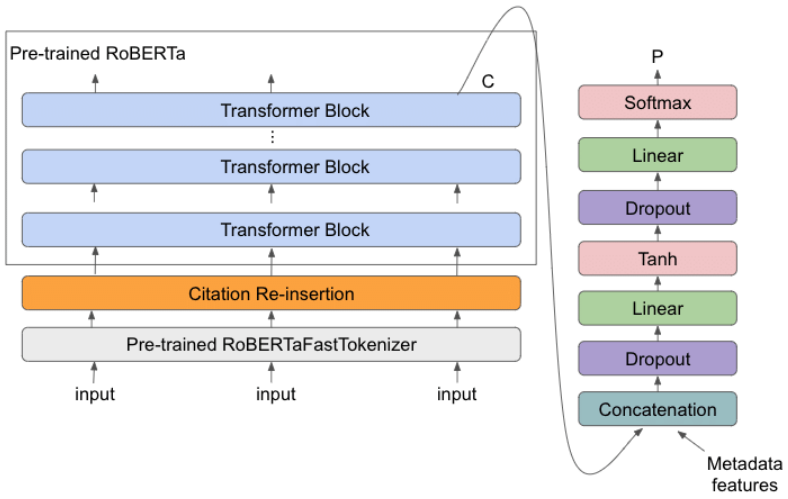

Figure 4 – RoBERTa model architecture (Huang et al. 2021)

[2] https://www.odbms.org/2023/02/sentiment-analysis-using-twitter-api-and-roberta-model/
[3] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

3.4 EDA – Exploratory Data Analysis

VADER Sentiment Analysis				
	Positive	Negative	Neutral	All Tweets
Average length of tweets	111	112	111	111
Average word count of tweets	16	16	16	16
Maximum length of tweets	227	247	253	253
Minimum length of tweets	4	3	1	1

RoBERTa Sentiment Analysis				
	Positive	Negative	Neutral	All Tweets
Average length of tweets	92	120	103	111
Average word count of tweets	14	17	15	16
Maximum length of tweets	222	253	243	253
Minimum length of tweets	4	4	1	1

Table 1 – Descriptive Statistics for VADER and RoBERTa

Table 1 demonstrates the aggregate number of tweets, average word count of tweets, maximum length of tweets, and minimum length of tweets of all the three sentiments for VADER and RoBERTa. The sentiment analysis performed by VADER and RoBERTa shows variations in tweet length and word count. RoBERTa has a shorter average length, a higher word count, and a wider range of lengths, showing detailed analysis. VADER has more consistent length but a lower word count.

3.4.1 VADER – EDA

Figure 5 – WordCloud of all Tweets

Figure 6 – WordCloud of Positive Tweets

Figure 7 – WordCloud of Negative Tweets

Figure 8 – WordCLoud of Neutral Tweets

As seen in figure 5, 6, 7 and 8, ‘russia’ and ‘ukraine’ are the most common words used in all the tweets. In addition to these, ‘putin’, ‘china’, ‘war’, and ‘nato’ are among the most common words used in positive and negative tweets. Along with these words, ‘trump’, and ‘people’ are few of the most used words in neutral tweets for VADER sentiment analysis.

3.4.2 RoBERTa – EDA

For RoBERTa sentiment analysis, along with ‘russia’, ‘russian’, and ‘ukraine’, words like ‘support’, and ‘great’ were most commonly used in positive tweets. In addition to this, ‘nato’, ‘china’, ‘putin’, and ‘country’ are among the most commonly used words in neutral and negative sentiment tweets (Refer figure 9, 10 and 11).

Figure 9 – WordCloud of Positive Tweets

Figure 10 – WordCloud of Negative Tweets

Figure 11 – WordCloud of Neutral Tweets

3.5 Models Used

3.5.1 BERT

BERT revolutionized natural language processing by reading text in both directions at the same time using Transformers. BERT advances over earlier models like RNN and CNN with its ability to analyse data in any order. The BERT model used in this project is BERT Base Uncased [4]. The model is Uncased, i.e., it does not differentiate between lowercase or uppercase words. For example, this model would treat Bert and bert in the same way. The model’s pre-training consisted of two tasks, namely, Masked Language Modelling (MLM) and Next Sentence Prediction (NSP). In MLM, the model randomly masks 15% of the words

[4] https://huggingface.co/bert-base-uncased

in the input and then passes the masked text through it to predict the masked words. In NSP, the model takes two masked sentences as inputs. These sentences may or may not be sequential in the original text. NSP helps the model to understand the context between sentences and helps to improve the model's knowledge of sequential language patterns (Devlin et al. 2018). The BERT tokenizer is important to convert the pre-processed tweets into numerical inputs for the model. Special tokens are added like, [CLS], [SEP], and [MASK] used for classification, sentence separation and masking during pre-training respectively. This tokenization and encoding technique help BERT understand the complex language structures and details in a better way. BERT tokenizer gives three components as outputs, namely, 'input_ids', the numerical identifiers of the vocabulary tokens, 'token_type_ids', identifies which part of the sentence each token belongs to, and 'attention_mask', informs the model about which tokens to prioritize and which to ignore, during training [5].

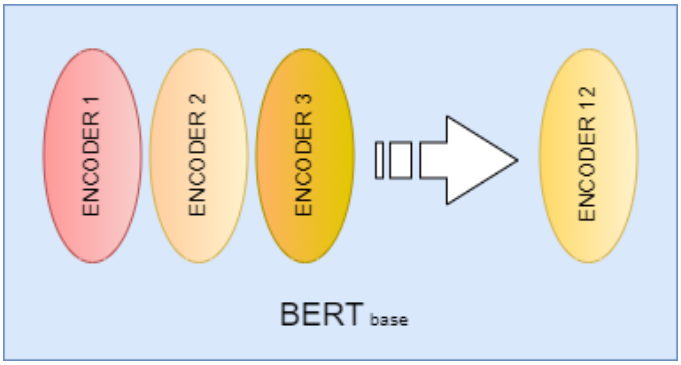

Figure 11 – BERT base model with 12 layers

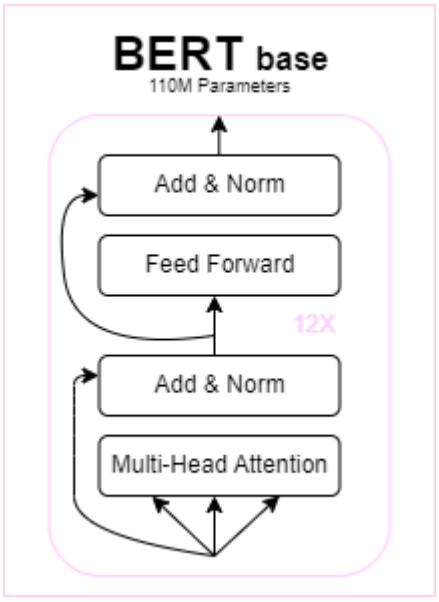

Figure 12 – BERT base model architecture

[5] https://www.geeksforgeeks.org/sentiment-classification-using-bert/

4 Implementation

In this project, two different approaches to label tweets dataset were considered. First approach is the Rule-based or the lexicon-based approach. In this approach VADER from the NLTK package was used. Lexicon-based models such as VADER, make use of dictionaries (lexicons) that contain words or emojis with positive or negative weights. The model evaluates the text by counting the occurrences of positive and negative words. If the positive word count is greater than the negative word count, the tweet is classified as positive and vice versa. The tweet is classified as neutral in the case where the count of positive and negative words is equal.

The second approach is the Unsupervised Deep learning approach. To perform sentiment analysis, models like RoBERTa, recognize patterns in the text by processing unstructured and unlabelled data over several layers and applying varied learning techniques such as self-attention.

Since it is not possible to compare the accuracies of these two approaches without the actual (true) sentiment for the tweets, the two labelled datasets were used – one labelled by VADER approach and the other labelled by RoBERTa approach, to fine-tune two similar BERT-base models and compare their accuracies.

The problem in dealing with unlabelled datasets and unsupervised learning is that true labels are not available for sentiment of the tweets to compare with the predicted labels and evaluate the performance of our model. In order to deal with this problem in the project, a unique approach is implemented where only those tweets for which the sentiment label was predicted same by the VADER and RoBERTa are filtered out. For example, if for a particular tweet, VADER and RoBERTa both predicts the sentiment label as 'Positive', then the true label of that tweet is considered to be positive. In this part of the project, the tweets that were labelled differently by the two approaches were discarded. Then the filtered data with true labels are fed to the BERT model in order to further train and fine-tune it (Refer figure 13).

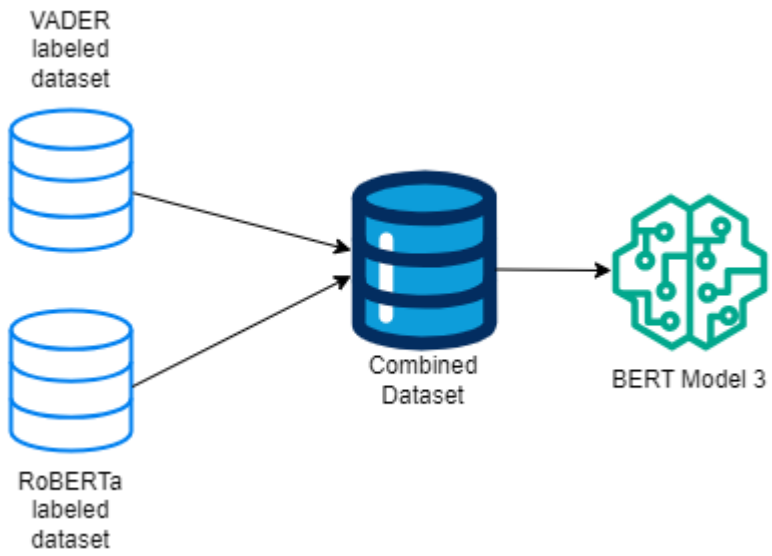

Figure 13 – Second Approach with True labels

The hyperparameters used in fine-tuning these BERT models are learning rate, batch size, and number of epochs. Learning rate determines the size of the step taken during the

optimization process. It impacts the learning rate of the model. A quicker learning rate allows for faster learning but increases the risk of exceeding the optimal weights, whereas a slower learning rate may converge slowly but with better precision. Batch size is the count of training examples utilized in one iteration. The dataset is split into batches during training, and the model's parameters are changed depending on the average loss across each batch. The batch size selected can affect the training speed and memory requirements. Training the model requires multiple iterations through the dataset, known as epochs. When the model has gone through each training example once, one epoch is completed. During training, the number of epochs defines how many times the model iterates over the complete dataset.

To attain consistent reproducibility, consistent seeds were set before data preprocessing, before data splitting, and before model initialization. Uniform data pre-processing steps were used in the same order for each model and before final results the entire code was re-run numerous times to ensure consistent results.

5 Evaluation

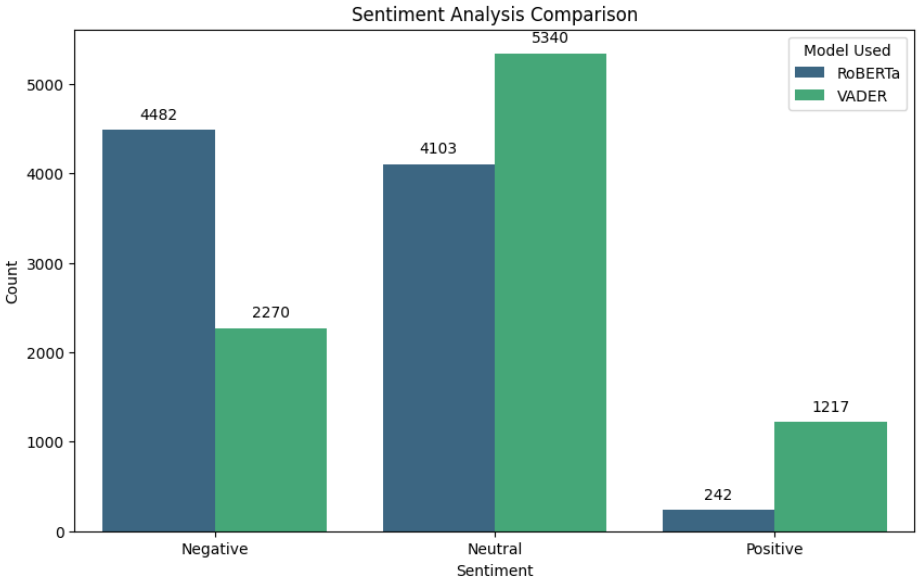

Figure 14 – Sentiment Analysis Comparison

After collection of data, cleaning of data, and selecting only English data, 8,827 tweets were obtained out of 10,014 tweets. Two approaches were implemented to label the tweets into one of the three sentiment categories – Positive, Negative, and Neutral. The two approaches used were VADER and RoBERTa. As shown in Figure 14, Negative is the most prevalent sentiment in the tweets labelled by RoBERTa, accounting for 38.1%, followed by Neutral, accounting for 34.9% and Positive, accounting for 2.1%. For VADER, the most prevalent sentiment is Neutral, accounting for 64.9%, followed by Negative, accounting for 27.6%, and Positive, accounting for 7.5% of the total tweets.

Two separate BERT models were trained and fine-tuned using the two labelled datasets. The models were trained on variety of learning rates such as 0.0001, 0.00001, and 0.000001, batch sizes of 16 and 32, and were run for different number of epochs. Fine tuning these parameter values helps to achieve best model performance and prevent overfitting and underfitting. The evaluation metrics chosen are accuracy, f1-score, precision, and recall. These evaluation metrics were chosen as they provide a comprehensive and balanced view of the model's performance across the three sentiment classes. Due to the presence of class imbalance in the data, accuracy might not adequately represent the model's effectiveness as it may be biased towards the majority class. F1-score is especially relevant where there is an imbalance between the number of instances belonging to different sentiment classes. Precision ensures that positive predictions are trustworthy while Recall ensures that the model does not miss important instances of positive sentiments [6].

To find the best optimum model, 18 different variations of VADER based BERT model and RoBERTa based BERT model were run, with different values of learning rates, batch sizes, and number of epochs to compare their accuracies (Refer Appendix 1, table 2 and 3). The best VADER based model achieved an average accuracy of 49.58% with the learning rate of 0.00001, batch-size of 16, and number of epochs as 10, while the best RoBERTa based model achieved an average accuracy of 83.017% with the learning rate of 0.00001, batch-size of 32, and number of epochs as 5 (Refer table 2). In terms of accuracy, the RoBERTa based model outperformed the VADER based model significantly, demonstrating that RoBERTa is a better fit for sentiment analysis/opinion mining on the given data. The values for learning rate, batch size and number of epochs were chosen using hit and trial method.

Best Metrics	**BERT with VADER**	**BERT with RoBERTa**
Average Accuracy	49.58	83.02
Average F1 Score	36.09	82.96
Average Precision	41.88	86.23
Average Recall	32.21	84.55

Table 2 – Model results

Figure 15 contains the boxplot for Accuracy, Precision, Recall, and F1-Score for the 10 re-runs of the best performing model for VADER based BERT model. Box plot of these metrics provides a visual summary of variation, central tendency, and potential outliers, helping in a

[6] https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/?right-decisions%2F#:~:text=Camon%20%20metrics%20%20include%20%20accuracy%20(proportion,curve%20(AUC%2DROC).

brief demonstration of performance distribution and stability. Similarly, figure 16 contains the boxplot for the same evaluation metrics for the re-runs of the best performing RoBERTa based model (Refer Appendix 1, table 4 and 5).

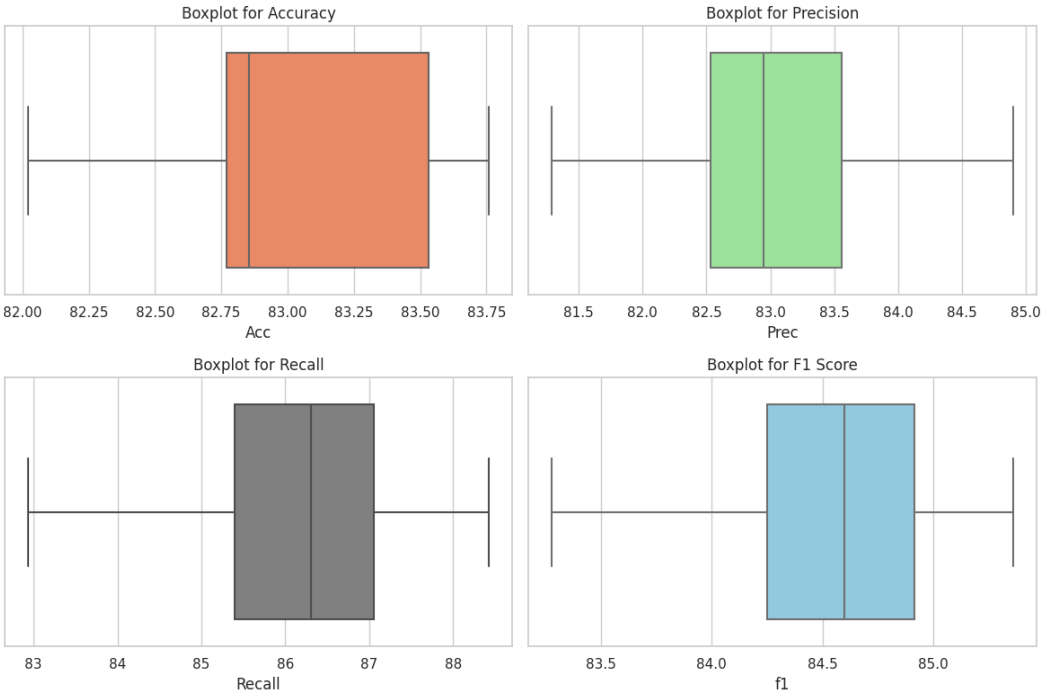

Figure 15 – Boxplot for Accuracy, Precision, Recall, and F1 Score for VADER based model

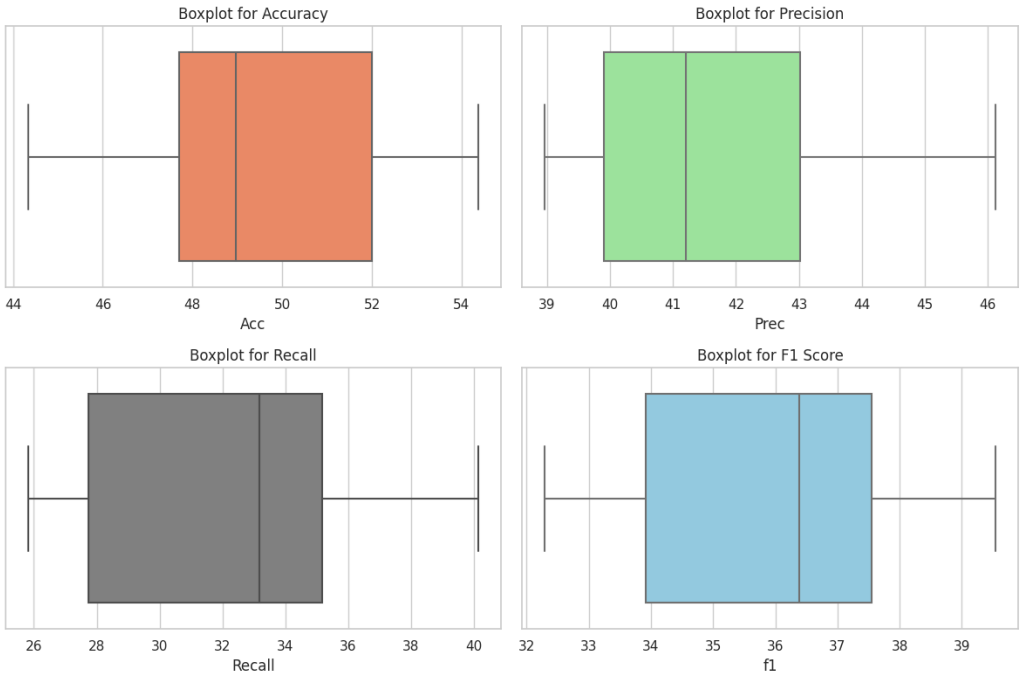

Figure 16 – Boxplot for Accuracy, Precision, Recall, and F1 Score for RoBERTa based model

Figure 17 shows the results of the re-runs performed for the best performing BERT model with true labels (approach 2). The best model performance was achieved with 0.00001 learning rate, batch size of 32, and number of epochs as 5. Table 3 encapsulates the average of metrics for the best performing model.

Best Metrics	Average Accuracy	Average Precision	Average Recall	Average F1-Score
BERT model (Approach 2)	83.95	76.89	72.56	74.47

Table 3 – Model results for Approach 2

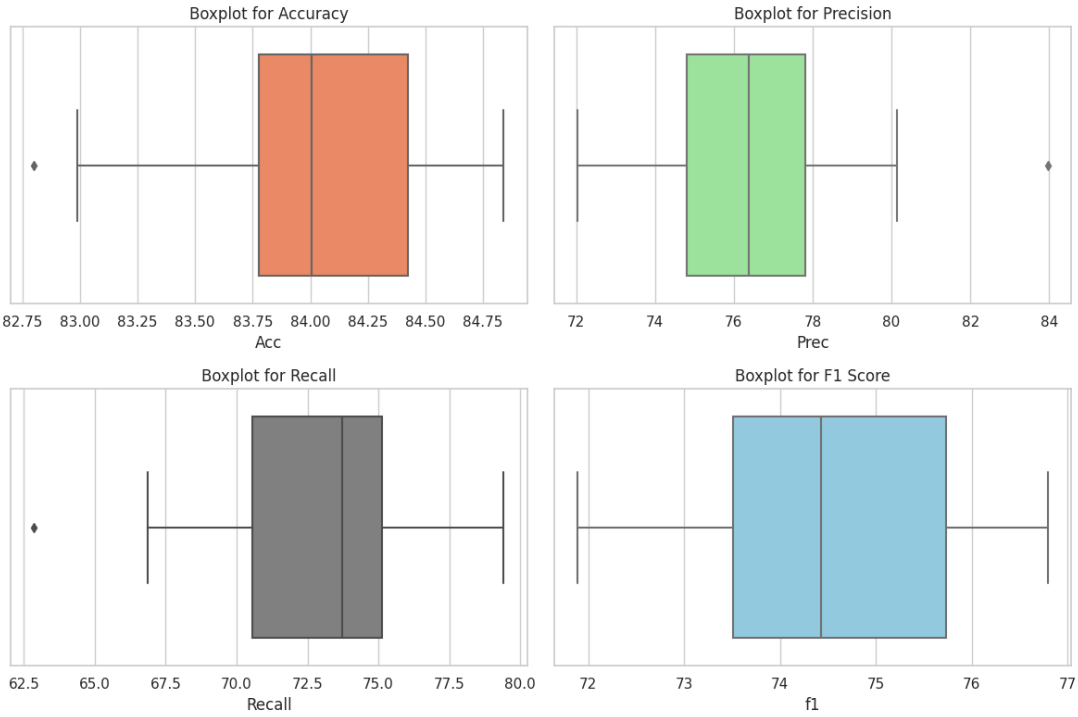

Figure 17 – Boxplot for Accuracy, Precision, Recall, and F1-Score for Approach 2

5.1 Discussion

Even though the full-scale invasion of Ukraine by Russia began on 24th February 2022, this geopolitical conflict has been in the headlines since 2014. The sentiment analysis results provided above are in line with the results depicted in the studies conducted by (Nandurkar et al. 2023), and (Ramos and Chang 2023). Most of the tweets in the dataset have either negative or neutral sentiment, and the count of positive sentiment tweets are very low. This shows that social media users, especially twitter users, are still against Russia's actions and are criticising Russian government's decision. These results in turn backs up the fact that both the models (VADER and RoBERTa), are performing adequately.

The discrepancy in the results obtained by VADER and RoBERTa are due to the fact that, VADER is a rule-based sentiment analysis tool, while RoBERTa is a transformer-based model. VADER uses a list of pre-defined words with associated sentiment scores to analyse the sentiment of the text. It takes into account the intensity of sentiments and incorporates grammatical standards. RoBERTa, on the other hand, leverages deep learning and contextual embeddings to understand the meaning of words in a sentence. RoBERTa does not rely on predefined sentiment scores but learns from contextual information in large datasets during training.

Due to class imbalance between positive, negative, and neutral tweets, the proposed BERT model might be biased towards the majority class. The model would have performed even better if there had been only a minor class imbalance, but as shown in figure 12, both the models – VADER and RoBERTa, have predicted more tweets as negative and neutral compared to positive. This class imbalance can be justified by the fact that the tweets collected for this project are related to Ukraine-Russia war, and generally people's sentiments are against crime, violence, and war, resulting in more negative and neutral tweets than positive tweets.

A few of the positive and neutral tweets would have been classified wrongly due to the presence of sarcasm or irony in the tweets. The state-of-the-art models consists of deep and complex architecture, which helps them to understand the meaning of the words in the sentence along with the context but fails to incorporate the presence of sarcasm in the text. For example, if the tweet is as follows – "Wow, Russia's 'peacekeeping efforts' in Ukraine are truly commendable. Their unique approach to territorial respect is quite impressive.", the use of words like 'Wow', 'commendable', 'quite impressive' may distract the model from capturing the sarcasm in the text and will label it as 'Positive' rather than 'Negative'.

This study's novel approach was to work on 2023 twitter tweets and compare the results to the 2022 studies. Even though the results of this study were in line with the studies from 2022, the unavailability of tweets spanning across several month is an issue that makes the results of this study less reliable. Since the dataset obtained only consists of the tweets from two days in February 2023, the results of this study cannot be generalised for the year 2023. Another unique approach was performed in this study, in which the labelled datasets of VADER and RoBERTa were merged to extract the true labels for the tweets. Out of the three fine-tuned BERT models, the models based on RoBERTa and the true labels performed well in comparison to the VADER based BERT model. The difference in the approach of VADER and RoBERTa to label tweets could be the reason for the poor performance of VADER and better performance of RoBERTa.

The best performing models were re-run several times to obtain the average of evaluation metrics such as accuracy, precision, recall, and f1-score. Despite using multiple set seeds across our project, subtle variations in these metrics can occur due to GPU utilization, uncontrolled hyperparameters, and insufficient tuning of hyperparameters. BERT base has 110 million hyperparameters including both trainable and non-trainable parameters due to which, slight variations may occur. To solve this issue, the average of evaluation metrics was achieved over 10 re-runs with same parameters.

Overall, this study was able to achieve good insights about the public sentiments relating to Ukraine-Russia war. The weights obtained during training of the BERT model were saved and can be implemented to predict the sentiment of a user-given text. With availability of

more data, incorporating the tweets from all the months of 2023, the fine-tuned BERT can be used to predict the sentiments in the tweets and give generalised insights about public opinion relating to Ukraine-Russia war in 2023.

6 Conclusion

This study analysed 8858 English tweets, aiming to understand and provide insights about the Ukraine-Russia War. This study was focused on getting insights about public sentiment in the year 2023 and compare it with the results of the studies conducted until 2022. The research question - "How can BERT models be effectively applied to perform opinion mining on a Twitter dataset related to the Ukraine-Russia war?" was successfully addressed. In addition to BERT (Bidirectional Encoder Representations from Transformers) and its variation, RoBERTa (Robustly optimised BERT approach), VADER (Valence Aware Dictionary and sEntiment Reasoner), which is a rule-based sentiment analysis tool instead of a transformer-based model, from the Natural Language Tool-Kit was also utilised and used for comparison. The evaluation results shows that the RoBERTa based BERT model outperformed VADER based BERT model with an accuracy difference of 33.44%.

The insights obtained from exploratory data analysis infers that, 'Negative' was the most common sentiment in RoBERTa (38.1%), while 'Neutral' was the most common sentiment in VADER (64.9%). The results from VADER and RoBERTa were merged in order to filter out tweets with true labels. The BERT model fine-tuned using this data gave an accuracy of 87.98%. This fine-tuned BERT can be utilized to predict the sentiment of user defined text. The sentiment analysis results indicate that, as anticipated, a high number of tweets related to the conflict were 'Negative', with negative sentiment mainly targeted towards Russia.

6.1 Future Work

- For further study and analysis, the results of this study can act as the base results and can be worked upon.
- In addition to accumulation of more diversified data for 2023, several other variations of BERT model specifically trained and fine-tuned for sentiment analysis task can be used with conjunction to the models used in this study for a comparative study.
- In addition to Positive, Negative, and Neutral sentiment, several other human emotions can be extracted from the tweets like, fear, joy, anger, etc. to enhance the study and delve deeper into the public sentiment and opinion mining relating to Ukraine-Russia war.
- Other suitable evaluation metrics such as confusion matrix and cross-entropy can be used to give a more detailed evaluation of the results.

References

Ahmed, S., Hasan, M.M. and Kamal, M.R. 2023. Russia–Ukraine crisis: The effects on the European stock market. *European Financial Management* 29(4), pp. 1078–1118. doi: 10.1111/eufm.12386.

Aslan, S. 2023. A deep learning-based sentiment analysis approach (MF-CNN-BILSTM) and topic modeling of tweets related to the Ukraine–Russia conflict. *Applied Soft Computing* 143. doi: 10.1016/j.asoc.2023.110404.

Barreto, S., Moura, R., Carvalho, J., Paes, A. and Plastino, A. 2023a. Sentiment analysis in tweets: an assessment study from classical to modern word representation models. *Data Mining and Knowledge Discovery* 37(1), pp. 318–380. doi: 10.1007/s10618-022-00853-0.

Barreto, S., Moura, R., Carvalho, J., Paes, A. and Plastino, A. 2023b. Sentiment analysis in tweets: an assessment study from classical to modern word representation models. *Data Mining and Knowledge Discovery* 37(1), pp. 318–380. doi: 10.1007/s10618-022-00853-0.

Bello, A., Ng, S.C. and Leung, M.F. 2023a. A BERT Framework to Sentiment Analysis of Tweets. *Sensors* 23(1). doi: 10.3390/s23010506.

Bello, A., Ng, S.C. and Leung, M.F. 2023b. A BERT Framework to Sentiment Analysis of Tweets. *Sensors* 23(1). doi: 10.3390/s23010506.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Available at: http://arxiv.org/abs/1810.04805.

Dominic, P., Purushothaman, N., Kumar, A.S.A., Prabagaran, A., Blessy, J.A. and John, A. 2023. Multilingual Sentiment Analysis using Deep-Learning Architectures. In: *Proceedings - 5th International Conference on Smart Systems and Inventive Technology, ICSSIT 2023*. Institute of Electrical and Electronics Engineers Inc., pp. 1077–1083. doi: 10.1109/ICSSIT55814.2023.10060993.

La Gatta, V., Wei, C., Luceri, L., Pierri, F. and Ferrara, E. 2023. Retrieving false claims on Twitter during the Russia-Ukraine conflict. In: *ACM Web Conference 2023 - Companion of the World Wide Web Conference, WWW 2023*. Association for Computing Machinery, Inc, pp. 1317–1323. doi: 10.1145/3543873.3587571.

Hasan, M., Islam, L., Jahan, I., Meem, S.M. and Rahman, R.M. 2023. Natural Language Processing and Sentiment Analysis on Bangla Social Media Comments on Russia-Ukraine War Using Transformers. *Vietnam Journal of Computer Science* 10(3), pp. 329–356. doi: 10.1142/S2196888823500021.

Huang, Z., Low, C., Teng, M., Zhang, H., Ho, D.E., Krass, M.S. and Grabmair, M. 2021. Context-aware legal citation recommendation using deep learning. In: *Proceedings of the 18th International Conference on Artificial Intelligence and Law, ICAIL 2021*. Association for Computing Machinery, Inc, pp. 79–88. doi: 10.1145/3462757.3466066.

Hutto, C.J. and Gilbert, E. 2014. *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. Available at: http://sentic.net/.

Li, J. 2023. Fine-Grained Sentiment Analysis with a Fine-Tuned BERT and an Improved Pre-Training BERT. In: *2023 IEEE International Conference on Image Processing and Computer Applications, ICIPCA 2023*. Institute of Electrical and Electronics Engineers Inc., pp. 1031–1034. doi: 10.1109/ICIPCA59209.2023.10257673.

Liu, Y. et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Available at: http://arxiv.org/abs/1907.11692.

Nandurkar, T., Nagare, S., Hake, S. and Chinnaiah, K. 2023. Sentiment Analysis Towards Russia - Ukrainian Conflict: Analysis of Comments on Reddit. In: *International Conference on Emerging Trends in Engineering and Technology, ICETET*. IEEE Computer Society. doi: 10.1109/ICETET-SIP58143.2023.10151571.

Ramos, L. and Chang, O. 2023. Sentiment Analysis of Russia-Ukraine Conflict Tweets Using RoBERTa. *Uniciencia* 37(1). doi: 10.15359/ru.37-1.23.

Sasmoko, Imran, M., Khan, S., Khan, H. ur R., Jambari, H., Musah, M.B. and Zaman, K. 2023. War psychology: The global carbon emissions impact of the Ukraine-Russia conflict. *Frontiers in Environmental Science* 11. doi: 10.3389/fenvs.2023.1065301.

Sheng, D. and Yuan, J. 2021. An Efficient Long Chinese Text Sentiment Analysis Method Using BERT-Based Models with BiGRU. In: *Proceedings of the 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2021*. Institute of Electrical and Electronics Engineers Inc., pp. 192–197. doi: 10.1109/CSCWD49262.2021.9437789.

Sirisha, U. and Chandana, B.S. [no date]. *Aspect based Sentiment and Emotion Analysis with ROBERTa, LSTM*. Available at: https://www.kaggle.com/code/ssaisuryateja/eda-.

Sohag, K., Islam, M.M., Tomas Žiković, I. and Mansour, H. 2023. Food inflation and geopolitical risks: analyzing European regions amid the Russia-Ukraine war. *British Food Journal* 125(7), pp. 2368–2391. doi: 10.1108/BFJ-09-2022-0793.

Thakkar, H., Patil, A., Saudagar, O. and Yenkikar, A. 2023. Sentiment and Statistical Analysis on Custom Twitter Dataset for 2022 Russo-Ukrainian Conflict. In: *Proceedings of the International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics, ICIITCEE 2023*. Institute of Electrical and Electronics Engineers Inc., pp. 679–684. doi: 10.1109/IITCEE57236.2023.10090995.

Vyas, P., Vyas, G. and Dhiman, G. 2023. RUemo—The Classification Framework for Russia-Ukraine War-Related Societal Emotions on Twitter through Machine Learning. *Algorithms* 16(2). doi: 10.3390/a16020069.

Wadhwani, G.K., Varshney, P.K., Gupta, A. and Kumar, S. 2023. Sentiment Analysis and Comprehensive Evaluation of Supervised Machine Learning Models Using Twitter Data on Russia–Ukraine War. *SN Computer Science* 4(4). doi: 10.1007/s42979-023-01790-5.

Appendix 1

1.1 Finding the best performing model for VADER based BERT model.

Learning Rate	Batch Size	Epochs	Accuracy
0.0001	16	3	13.82
0.0001	16	5	60.40
0.0001	16	10	25.67
0.0001	32	3	13.82
0.0001	32	5	60.49
0.0001	32	10	13.82
0.00001	16	3	36.48
0.00001	16	5	32.70
0.00001	**16**	**10**	**54.38**
0.00001	32	3	28.70
0.00001	32	5	35.95
0.00001	32	10	25.52
0.000001	16	3	33.53
0.000001	16	5	33.30
0.000001	16	10	35.50
0.000001	32	3	35.57
0.000001	32	5	27.87
0.000001	32	10	30.66

Table 1 – VADER based BERT model performance

1.2 Finding the best performing model for RoBERTa based BERT model.

Learning Rate	Batch Size	Epochs	Accuracy
0.0001	16	5	76.81
0.0001	16	10	76.05
0.0001	16	20	50.75
0.0001	32	5	78.02
0.0001	32	10	78.24
0.0001	32	20	77.41
0.00001	16	5	82.77
0.00001	16	10	83.00
0.00001	16	20	80.89
0.00001	**32**	**5**	**83.76**
0.00001	32	10	82.85
0.00001	32	20	81.26
0.000001	16	5	79.45
0.000001	16	10	80.74
0.000001	16	20	82.85
0.000001	32	5	76.81
0.000001	32	10	80.96

0.000001	32	20	82.02

Table 2 – RoBERTa based BERT model performance

1.3 Finding the best performing model for the BERT model trained using True labeled dataset (Approach 2).

Learning Rate	Batch Size	Epochs	Accuracy
0.0001	16	3	79.66
0.0001	16	5	80.03
0.0001	16	10	67.65
0.0001	32	3	79.11
0.0001	32	5	79.66
0.0001	32	10	67.65
0.00001	16	3	83.91
0.00001	16	5	85.21
0.00001	16	10	86.50
0.00001	32	3	83.17
0.00001	**32**	**5**	**87.98**
0.00001	32	10	86.69
0.000001	16	3	70.79
0.000001	16	5	77.81
0.000001	16	10	83.17
0.000001	32	3	68.40
0.000001	32	5	70.79
0.000001	32	10	79.66

Table 3 – Approach 2 - BERT model performance

1.4 Re-runs of the best performing model for VADER based BERT model.

Accuracy	Precision	Recall	F1 Score
44.33	38.96	40.15	39.54
52.41	43.13	25.81	32.29
50.75	40.7	28.87	33.78
48.03	41.7	37.47	39.47
46.67	39.82	35.56	37.57
53.7	46.12	27.34	34.33
54.38	**45.87**	**26.57**	**33.65**
49.84	42.68	33.46	37.51
47.58	39.91	34.03	36.73
48.11	39.9	32.88	36.05

Table 4 – best VADER based BERT model performance

1.5 Re-runs of the best performing model for RoBERTa based BERT model.

Accuracy	Precision	Recall	F1 Score
82.78	81.29	88.29	84.65
82.47	81.78	86.74	84.18
82.93	82.72	86.45	84.55
82.77	83.31	85.19	84.23
82.77	82.65	86.03	84.31
83.61	82.49	88.43	85.36
82.02	83.64	82.93	83.28
83.76	**84.90**	**84.90**	**84.90**
83.53	83.17	87.16	85.12
83.53	83.69	86.17	84.92

Table 5 – best RoBERTa based BERT model performance

1.6 Re-runs of the best performing model for the BERT model – Approach 2.

Accuracy	Precision	Recall	F1 Score
83.91	80.13	66.85	72.89
83.73	77.7	69.71	73.49
84.84	77.84	74.28	76.02
84.1	83.96	62.85	71.89
84.47	76.3	75.42	75.86
82.8	72.04	76.57	74.23
84.47	74.33	79.42	76.79
82.99	73.99	73.14	73.56
83.91	76.19	73.14	74.63
84.28	76.47	74.28	75.36

Table 6 – best Approach 2 BERT model performance