Integrated Sentiment and Emotion Analysis of the Ukraine-Russia Conflict Using Machine Learning and Transformer Models

Mohammad Hossein Amirhosseini*, Nabeela Berardinelli, Kunal Gaikwad,
Christian Eze Iwuchukwu and Mahmud Ahmed
University of East London, London, U.K.

Keywords: Natural Language Processing, Sentiment Analysis, Emotion Analysis, Machine Learning, Social Media Analysis, Ukraine-Russia Conflict.

Abstract: The Russia-Ukraine war has been a significant international conflict, generating a wide range of public sentiments. With escalating geopolitical tensions, determining whether public discourse supports or condemns the invasion has become increasingly important. This study investigates public attitudes through large-scale sentiment analysis of 1,426,310 tweets collected during the early phase of the conflict. Sentiment classification was performed using machine learning models, including XGBoost, Random Forest, Naïve Bayes, Support Vector Machine, and a Feedforward Deep Learning model, combined with Count Vectorizer and TF-IDF. The deep learning model with Count Vectorizer achieved the highest accuracy at 89.58%, outperforming all others. To go beyond polarity classification, emotion prediction was also conducted using a lexicon-based method (NRC Emotion Lexicon) and a transformer-based model (DistilRoBERTa), both trained to classify tweets into eight emotions: joy, trust, surprise, fear, anger, sadness, disgust, and anticipation. A comparative evaluation showed that the transformer model significantly outperformed the lexicon-based model across all metrics, including accuracy, precision, recall, F1 score, and Hamming loss. Fear and anger emerged as the most dominant emotions, highlighting widespread public anxiety and distress. This analysis provides a nuanced understanding of online discourse during conflict and offers insights for researchers, policymakers, and communicators responding to global crises.

1 INTRODUCTION

The Russia-Ukraine conflict has been a significant international issue since 2014, causing geopolitical tensions, economic sanctions, and military action that led to a full-scale war in 2022. The conflict has generated various opinions among people worldwide. When it comes to how the media affects public opinion on the conflict between Russia and Ukraine, news organisations and social media have been particularly influential. The conflict has been portrayed differently in the media across nations and platforms, which has widened public opinion gaps. As high death rates have been demonstrated to negatively affect public opinion about a conflict, the influence of the conflict's death rate has had a substantial impact on attitudes around the war. International ties have been strained because of the conflict's toll on human life, which has influenced opinions of the war and its effects. Furthermore, the impact of financial assistance from Western nations has also shaped public attitudes surrounding the conflict.

With the escalation of conflict, it has become more crucial to determine whether the universal message about this conflict is an affirmation or a condemnation of the invasion. Social media platforms have emerged as a significant source for tracking public sentiment. Numerous users of social networks publish countless posts expressing their perspectives and emotions surrounding global events. Twitter, in particular, has served as a primary medium for real-time public discourse. In March 2022, the conflict between Ukraine and Russia was the most frequently tweeted topic (Al Maruf et al., 2022). Numerous tweets with the *#StandwithUkraine* hashtag expressed support for Ukraine and disapproval of Russia's actions during the conflict (Baker and Taher, 2023). Thus, references to the countries, their populations, and their respective administrations were intertwined with rejections of the conflict. This predominance is reflected in the keyword analysis of

Amirhosseini, M. H., Berardinelli, N., Gaikwad, K., Iwuchukwu, C. E., Ahmed and M.
Integrated Sentiment and Emotion Analysis of the Ukraine-Russia Conflict Using Machine Learning and Transformer Models.
DOI: 10.5220/0013645500003967
In *Proceedings of the 14th International Conference on Data Science, Technology and Applications (DATA 2025)*, pages 191-202
ISBN: 978-989-758-758-0; ISSN: 2184-285X

the content disseminated by the audience of the Office of the President of Ukraine's profile (Baker and Taher, 2023).

Understanding public sentiment in such contexts is critical—not only for researchers and policymakers but also for the media, humanitarian agencies, and the broader public. Natural Language Processing (NLP) is a computational method that utilizes various theories and technologies to interpret and analyse human language (Amirhosseini et al. 2018). Sentiment analysis, a technique grounded in Natural Language Processing and machine learning, allows us to computationally detect whether expressions in text are positive, negative, or neutral (Zhang et al., 2018; Dang et al., 2020; Stine 2019). This method has been widely used across domains including product reviews, political discourse, and crisis communication. While prior research has often relied on traditional models like Naïve Bayes or Support Vector Machine (Ahmad et al., 2017; Baid et al., 2017; Hasan et al., 2018; Jagdale et al., 2019; Bhavitha et al., 2017), the emergence of more powerful classifiers such as XGBoost, Random Forest, and deep learning architectures presents an opportunity to improve performance in large-scale sentiment analysis tasks.

In this study, we examine Twitter data from the early months of the Russia-Ukraine conflict to analyse the public’s attitude using five machine learning models. We employ both Count Vectorization and TF-IDF for feature extraction and compare model performance based on key classification metrics.

However, we recognise that binary or ternary sentiment classification (positive, negative, neutral) may not fully capture the emotional complexity expressed in crisis-related content. To address this limitation, this study also incorporates emotion detection, another subset of Natural Language Processing, which identifies specific affective states such as fear, anger, trust, joy, and sadness. Emotions offer a deeper lens through which to understand public discourse, as they shape political attitudes, influence behaviour, and reflect psychological responses to conflict (Mohammad and Turney, 2013; Sailunaz and Alhajj, 2019). Using both lexicon-based methods (NRC Emotion Lexicon) and transformer-based models (DistilRoBERTa), we mapped out eight core emotions and compared their performance in capturing the emotional framing of the war.

By combining sentiment analysis with multi-label emotion classification, this study contributes to a more nuanced understanding of how people emotionally engage with global conflicts online. In doing so, it offers important insights for researchers in NLP, social sciences, and political communication, while informing decision-making processes in policy and public diplomacy.

2 LITERATURE REVIEW

2.1 The Concept of Sentiment Analysis

Sentiment analysis is a research technique that involves mining user opinions on social media to understand attitudes toward services, products, politics, and events (Zhang et al., 2018). Defined as the computational analysis of sentiments, perspectives, and emotions, this method has grown with the rise of social media, which offers users a platform to express opinions (Zhang et al., 2018). Businesses and researchers use this data to understand public perception and behaviour.

According to Dang et al. (2020), sentiment analysis focuses on collecting and analysing sentiments shared online, particularly in the Web 2.0 era, where users frequently express views on diverse topics. It is a statistical approach that identifies patterns and trends from user-generated content (Stine, 2019; Dang et al., 2020; Zhang et al., 2018).

Stine (2019) explains that sentiment analysis categorizes text into positive or negative sentiments, aiming to summarise public opinion. This positivist approach helps extrapolate general attitudes from textual data.

There are three primary methods used in sentiment analysis: (1) machine learning-based approaches, (2) rule-based systems, and (3) lexicon-based models. These will be explored in the following sections.

2.2 Machine Learning-Based Sentiment Analysis

The widespread use of platforms like Twitter and Facebook has led to massive amounts of user-generated content that require automated methods for effective sentiment analysis. Ahmad et al. (2017) argue that manual analysis is infeasible at this scale, making machine learning essential for processing large datasets.

Machine learning techniques are well-suited for sentiment analysis, enabling the classification of user opinions through models that process Unigrams, Bigrams, and N-grams (Ahmad et al., 2017). These models typically perform binary classification to predict whether sentiments are positive or negative, which is applicable for analysing public attitudes toward events such as the Russia-Ukraine conflict.

Ahmad et al. (2017) and Baid et al. (2017) highlight the effectiveness of models like Naïve Bayes and K-Nearest Neighbour for sentiment classification. Naïve Bayes, known for its simplicity and scalability, assumes feature independence—an assumption that can limit its applicability when cultural or contextual factors influence sentiment.

Hasan et al. (2018) support this view, noting that while Naïve Bayes is useful for classifying Twitter sentiments, it provides limited insight into underlying causes. Jagdale et al. (2019) found similar results when applying Naïve Bayes and Support Vector Machine (SVM) to Amazon camera reviews. Bhavitha et al. (2017) add that Naïve Bayes performs well on small feature sets, while SVM excels with larger ones.

Overall, while Naïve Bayes offers fast and accurate results for small datasets, more robust techniques like SVM or Random Forest may be better suited for large-scale analysis. The choice of algorithm should therefore depend on dataset size and complexity.

2.3 Rule-Based Sentiment Analysis

Rule-based sentiment analysis relies on predefined linguistic rules to classify sentiments, often depending heavily on grammatical correctness (Ray & Chakrabarti, 2020). This makes it less effective when dealing with informal or unstructured language, which is common on social media. Vashishtha and Susan (2019) emphasize that grammatical accuracy is essential for this approach, and their study combined rule-based methods with deep learning to improve aspect-level sentiment classification by rephrasing informal expressions into grammatically correct forms.

While rephrasing can be time-consuming for large datasets, the rule-based approach has notable advantages. Dwivedi et al. (2019) argue that it offers simplicity, interpretability, and precision without requiring advanced computational resources. Berka (2020) adds that its ease of use and independence from training datasets make it practical for applications involving familiar languages.

Rule-based methods are often combined with lexicons to improve performance. Asghar et al. (2017) used a set of predefined "if-then" rules to classify emotion indicators like slang and emotion-specific terms, enabling phrase-level categorization of sentiment. Chekima et al. (2017) expanded on this by incorporating structured rules—such as intra-clause and extra-sentence patterns—and a term-counting strategy to detect polarity shifts, surpassing basic keyword-matching techniques.

In summary, despite limitations with informal language and implicit sentiment, rule-based models remain effective for clear, structured texts and serve as a valuable tool when interpretability and domain-specific knowledge are essential.

2.4 Lexicon-Based Sentiment Analysis

Lexicon-based sentiment analysis is one of the core approaches for automatically categorizing opinions and emotions in text. According to Khoo and Johnkhan (2018), unlike machine learning models that rely on supervised training and feature vectors such as unigrams or n-grams, the lexicon-based method uses predefined word lists annotated with sentiment polarity—positive, negative, or neutral.

Bonta and Janardhan (2019) define lexicon-based analysis as the classification of words or phrases based on their semantic orientation using a sentiment lexicon. These lexicons are typically developed through corpus-based, manual, or lexical methods. Polarity scores are then assigned to text based on the frequency and intensity of matched sentiment-bearing terms. Aung and Myo (2017) employed this approach to assess student feedback, using a curated database of opinion words with corresponding sentiment scores (see Table 1). The lexicon included not only adjectives and verbs but also intensifiers that

Table 1: Sentiment word database (Aung and Myo, 2017).

Example Opinion Words		
Opinion Word	***Score***	***Description***
Care	+2	Verb
Useful	+2	Adjective
Helpful	+2	Adjective
Clear	+2	Adjective + Verb
Good	+2	Adjective + Verb
Joyful	+1	Adjective
Marvellous	+3	Adjective
Brilliant	+3	Adjective
Ordinary	0	Adjective
Complex	-3	Adjective
Confuse	-3	Verb
Normal	0	Adjective
Complicated	-3	Adjective
Sleepy	-2	Adjective
Fast	-1	Adjective
Daily	0	Adjective
most	+100%	Intensifier
slightly	-50%	Intensifier
really	+25%	Intensifier
little	-50%	Intensifier
very	+50%	Intensifier
easily	+25%	Intensifier

modify sentiment strength.

While the lexicon model is computationally efficient and easy to implement, it requires domain-specific knowledge to build accurate word lists. Aung and Myo (2017) and Dehghani et al. (2017) note that the manual construction of such databases can be labour-intensive and requires programming expertise. Nevertheless, this approach remains useful for identifying general attitudes on various topics when contextual nuance and sarcasm are limited.

2.5 Challenges and Limitations of Sentiment Analysis

Sentiment analysis faces several challenges, particularly when applied to social media data. One major issue is domain dependence—words that are positive in one context may not be so in another. For example, a term considered positive in hospitality may carry a different connotation in education (Hussein, 2018). Both machine learning and lexicon-based approaches can struggle with this limitation, potentially leading to inaccurate sentiment classification if domain-specific nuances are overlooked (Aung and Myo, 2017). To address this, manually curated sentiment lexicons tailored to specific domains are often necessary.

Another challenge involves annotation and labeling. Mohammad (2017) notes that sentiment labeling often relies on human intuition, which can introduce inconsistency, especially when dealing with complex or ambiguous expressions. Aung and Myo (2017) similarly highlight the difficulty of developing reliable annotation schemes without clear contextual cues.

Subjectivity is also a key concern. According to Chaturvedi et al. (2018), the annotator's personal knowledge, experience, and interpretation can influence sentiment classification. This subjectivity complicates both manual labeling and model training, making it critical to implement strategies that reduce bias and ensure consistency in sentiment analysis.

2.6 Sentiment Analysis of Twitter Data

Twitter provides a rich source of real-time public opinion, and sentiment analysis enables researchers to extract insights from this content. Ahuja and Dubey (2017) explored sentiment analysis techniques for Twitter, using clustering methods to group tweets by sentiment polarity. Their study showed that tweets could be effectively classified using multiple dictionaries and that clustering helped distinguish varying degrees of positive and negative sentiment. Wang et al. (2018) supported this approach, highlighting how word clustering enhances classification accuracy. They introduced a Chi Square-based feature clustering and weighting technique, which improves sentiment detection when paired with models like Naïve Bayes.

Clustering is also useful within lexicon-based sentiment analysis. Mostafa (2019) applied it to categorize Twitter users' sentiments on halal food into four distinct groups, demonstrating its utility beyond machine learning contexts.

In addition to clustering, deep learning techniques have shown promise in analysing Twitter data. Liao et al. (2017) used convolutional neural networks (CNNs) to mine Twitter sentiment and found CNNs to be more effective than traditional methods like Naïve Bayes. Similarly, Huq et al. (2017) found both deep learning and classic classifiers like SVM and K-Nearest Neighbour capable of yielding accurate sentiment labels. As these studies suggest, the choice of classification method significantly influences the effectiveness of sentiment analysis on Twitter.

2.7 Emotion Detection in Social Media Analysis

While sentiment analysis classifies opinions into broad categories like positive, negative, or neutral, it often fails to capture the emotional depth found in social media—especially during crises like the Russia-Ukraine war. To address this, researchers have increasingly adopted emotion detection, which focuses on identifying emotions such as fear, anger, joy, and trust (Alhindi et al., 2018; Mohammad and Turney, 2013).

Emotion detection allows for a richer understanding of public sentiment by uncovering the psychological and affective states behind expressions (Sailunaz and Alhajj, 2019). This is particularly relevant on platforms like Twitter, where users frequently react to global events with emotionally charged content. A widely used tool in lexicon-based emotion detection is the NRC Emotion Lexicon (Mohammad and Turney, 2013), which categorizes words into eight core emotions from Plutchik's model: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust. Although interpretable and efficient, lexicon-based methods struggle with sarcasm and contextual subtleties common in social media (Yadollahi et al., 2017).

To overcome these limitations, transformer-based models like DistilRoBERTa and BERT have been adopted for emotion classification due to their ability to capture semantic nuance (Hartmann et al., 2022;

Liu et al., 2019). Hartmann's DistilRoBERTa, fine-tuned for English emotion detection, outperforms lexicon models in recognizing mixed or ambiguous emotions in short texts.

Comparative studies confirm that transformer models surpass lexicon approaches in precision and recall, particularly for multilabel classification (Gupta et al., 2022). Despite these developments, few studies have evaluated both approaches side-by-side in the context of the Russia-Ukraine conflict. This study addresses that gap by integrating NRC Emotion Lexicon (Mohammad and Turney, 2013) and DistilRoBERTa (Hartmann et al., 2022), providing a comparative analysis that enhances the validity and interpretability of emotion detection in geopolitical discourse.

2.8 Research Gap

The literature shows that sentiment analysis is highly domain-dependent, requiring separate investigation for each context to accurately capture user emotions. While many studies focus on products or politics, few explore real-time emotional responses to global conflicts like the Russia-Ukraine war.

Most research on the Russia-Ukraine conflict focuses on binary sentiment, with little attention to specific emotions like fear, anger, sadness, and hope. Capturing these emotions offers deeper insights into public narratives, but tools like the NRC lexicon often miss the subtleties of informal social media language.

Moreover, while traditional models like Naïve Bayes and SVM are common in sentiment analysis, newer models such as Random Forest, XGBoost, and Feedforward deep learning have been less explored in this context, despite their potential to better capture non-linear patterns and handle large-scale, high-dimensional data.

This study addresses this gap by implementing and evaluating these advanced models in comparison with baseline techniques, showing that deep learning and XGBoost models significantly outperform traditional classifiers in both accuracy and robustness.

To complement sentiment analysis, this study introduces emotion detection using both lexicon-based (NRC Emotion Lexicon) and transformer-based (DistilRoBERTa) approaches. Although emotion analysis is growing in NLP, its use on large-scale conflict-related social media data remains limited. This paper pioneers such integration for the Russia-Ukraine conflict, offering a comparative evaluation of the two methods.

In summary, the main research gaps addressed in this study are:

- A lack of sentiment and emotion analysis focused specifically on the Russia-Ukraine war using a large-scale, real-time Twitter dataset.
- Limited empirical comparison of traditional ML models with modern architectures like XGBoost and Feedforward deep learning in this context.
- Under exploration of emotion-specific classification using both lexicon-based and transformer models for richer interpretability.
- Absence of a combined sentiment–emotion analysis framework that can help policymakers understand not only public stance but also emotional framing.

By addressing these gaps, this study makes a unique contribution to both the sentiment analysis literature and the broader field of computational social science.

3 METHODOLOGY

3.1 Data Collection

Twitter API was used to collect data related to the conflict between Russia and Ukraine. Search keywords included: (1) "ukraine war", (2) "ukraine troops", (3) "ukraine border", (4) "ukraine NATO", (5) "StandwithUkraine", (6) "russian troops", (7) "russian border ukraine", and (8) "russia invade". Tweets were collected from 1st January 2022 to 6th March 2022. A maximum of 5,000 tweets were retrieved per day and stored in separate CSV files, which were later merged into a final dataset of 1,426,310 tweets.

3.2 Ethical Considerations

Ethical standards were upheld by ensuring that all collected data came from legitimate, publicly available sources and was used responsibly. The study did not involve direct data collection from individuals, minimizing concerns about privacy and confidentiality. In line with Twitter's data use policy, no data was sold or misused. The data was solely used for academic purposes and was anonymised to prevent user identification. As noted by Zimmer and Proferes (2014), conclusions were drawn strictly from the data without unsupported assumptions.

3.3 Data Cleaning

During the data cleaning and preprocessing steps, all the duplicated tweets were removed from the dataset.

As mentioned earlier, the original dataset included 1,426,310 tweets and after removing duplicated tweets, 1,313,818 tweets remained. We also realised that only 1,204,218 tweets are in English language. Thus, non-English tweets were removed from the dataset. Following these steps, missing values were dropped and user tags starting with '@'. URLs were also removed from the dataset.

3.4 Sentiment Analysis and Labelling Process

VADER (Valence Aware Dictionary and sEntiment Reasoner), a sentiment analysis tool from the NLTK library, was used to evaluate the sentiment of the tweets. This method combines lexicon-based and rule-based strategies, using a set of pre-labelled lexical features—words identified as conveying positive or negative sentiment—to classify new text accordingly (Bhatt et al., 2023). For each tweet, VADER computes sentiment scores for negativity, positivity, and neutrality using the *SentimentIntensityAnalyzer* module. Based on these scores, each tweet was assigned a sentiment label including Negative, Positive, or Neutral.

3.5 Preprocessing Steps

The dataset was split into features (X) and labels (y), with ‘content’ as the text input and ‘Sentiment’ as the target label. Preprocessing included lowercasing, emoji removal, tokenization, stop word removal, lemmatization, and punctuation removal—standard steps recommended in the literature for text data. Sentiment labels were numerically encoded for model compatibility. To convert text into numerical features, both Count Vectorization and TF-IDF were used to enable comparative performance analysis. Feature selection was applied using *SelectPercentile*, followed by standard scaling to ensure unit variance. The data was then stratified into training (80%) and test (20%) sets for model evaluation.

3.6 Implementation of the Machine Learning Models

In this study, five machine learning models were implemented for sentiment prediction: (1) XGBoost, (2) Random Forest, (3) Naïve Bayes, (4) Support Vector Machine, and (5) a Feedforward Deep Learning model. The models were developed using the Scikit-learn library and TensorFlow. GridSearchCV was employed for hyperparameter tuning of each model. Table 2 presents the optimal parameter values selected for each.

Table 2: Hyperparameter tuning outcomes.

Model	Parameter	Value
XGBoost	'n_estimators'	200
	'max_depth'	5
	'learning_rate'	0.1
	'subsample'	1.0
Random Forest	'n_estimators'	100
	'max_depth'	20
	'min_samples_split'	5
	'min_samples_leaf'	2
	'bootstrap'	True
Naïve Bayes	‘estimator’	MultinomialNB
	‘param_distributions’	parameters_nb
	‘n_iter’	5
	‘cv’	5
	‘verbose’	2
	‘random_state’	42
	‘n_jobs’	-1
Support Vector Machine	‘C’	1.0
	‘kernel’	rbf
	‘degree’	3
	‘random_state’	42
	‘tol’	0.001
	‘cache_size’	200
	‘max_iter’	1
Deep Learning	‘learning_rate’	0.0004
	‘num_hidden_layers’	5
	‘num_neurons_layer_0’	235
	‘num_neurons_layer_1’	237
	‘num_neurons_layer_2’	45
	'num_neurons_layer_3'	87
	'num_neurons_layer_4'	50
	‘dropout_rate’	0.34677

3.7 Emotion Detection and Using Lexicon and Transformer Models

In addition to sentiment classification, this study investigated emotional expressions in tweets to uncover deeper psychological and social narratives embedded within public discourse. Two distinct emotion detection approaches were employed. First, the NRC Emotion Lexicon (Mohammad & Turney, 2013) was used to map tokens within each tweet to eight core emotions: joy, trust, surprise, fear, anger, sadness, disgust, and anticipation.

Emotion scores were derived based on the frequency of emotionally associated words, and visualizations—such as word clouds and emotion distribution plots—were generated to highlight dominant emotional trends across the dataset.

To complement this lexicon-based method, we also applied a transformer-based model (j-hartmann/emotion-english-distilroberta-base). This fine-tuned, lightweight version of DistilRoBERTa is specifically trained for multilabel emotion classification in English-language social media text. It was used to classify tweets into the same eight emotion categories, allowing for a more context-sensitive interpretation of emotional content.

A comparative evaluation between the two models was conducted using a manually labeled benchmark set, focusing on micro-averaged F1 scores and other standard metrics to assess predictive performance.

4 RESULTS AND DISCUSSION

4.1 Sentiment Distribution in Tweets

The initial sentiment analysis was conducted using the VADER sentiment analysis tool, which classified tweets as Negative, Positive, or Neutral based on computed polarity scores. This step allowed for a high-level categorization of public opinion related to the Russia-Ukraine conflict. Out of the 1,204,218 English-language tweets included in the dataset, 32% were labelled as Negative, 20% as Positive, and 48% as Neutral. These findings, visualized in Figure 1, suggest a predominant leaning toward negative sentiment in the public discourse during the early stages of the war. The relatively high proportion of neutral tweets may reflect a cautious or observational tone among users, possibly indicative of a tendency to share factual updates or refrain from overt emotional engagement during a highly sensitive geopolitical event.

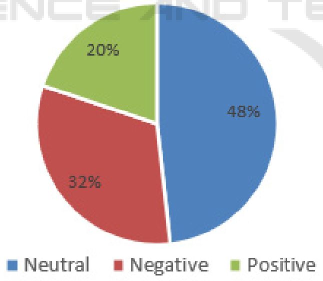

Figure 1: Percentage of the 'Negative,' 'Positive,' or 'Neutral' tweets.

4.2 Model Performance for Sentiment Prediction

To evaluate sentiment prediction performance, five machine learning models were implemented and trained: XGBoost, Random Forest, Naïve Bayes, Support Vector Machine, and a Feedforward Deep Learning model. Both Count Vectorizer and TF-IDF vectorization techniques were applied to transform the text data into numerical features, allowing for a comparative analysis of feature representation methods. Each model underwent 5-fold cross-validation to ensure robustness and generalizability of results. Performance was evaluated based on four key metrics: accuracy, precision, recall, and F1 score. The outcomes of this evaluation are presented in Table 3.

Among all models, the Feedforward Deep Learning model using Count Vectorization demonstrated the highest performance, achieving an accuracy of 89.58%, with correspondingly strong values for precision, recall, and F1 score. XGBoost followed closely with an accuracy of 87.01%, while the Random Forest model performed reasonably well with an accuracy of 83.40%. The Support Vector Machine and Naïve Bayes models yielded comparatively weaker results, with lower scores across all metrics. Furthermore, the Count Vectorizer generally led to better model performance than TF-IDF across all classifiers, likely due to its simplicity and effectiveness in capturing term frequency patterns within informal and short-form text like tweets.

Count Vectorizer outperformed TF-IDF likely because, in short tweets, frequent emotional words are key for sentiment detection, and TF-IDF down-weights these important terms while Count Vectorizer preserves them.

4.3 Emotion Analysis of Twitter Discourse

Beyond polarity sentiment classification, this study employed two distinct methodologies to examine the emotional dimension of public discourse: the NRC Emotion Lexicon, a rule-based system categorizing words into eight basic emotions, and a transformer-based model (DistilRoBERTa), fine-tuned for multi-label emotion classification on English social media content.

Table 3: 5-Fold cross validation results.

Model	Accuracy		Precision		Recall		F1 Score	
	Count Vec	*TF-IDF Vec*	*Count Vec*	*TF-IDF Vec*	*Count Vec*	*TF-IDF Vec*	*Count Vec*	*TF-IDF Vec*
XGB	87.01	86.15	86.80	86.05	87.01	86.15	86.56	85.80
RF	83.40	82.38	84.56	84.18	83.40	82.38	82.92	81.86
DL	**89.58**	**88.41**	**89.56**	**88.35**	**89.58**	**88.41**	**89.55**	**88.35**
SVM	77.91	77.83	77.85	77.66	77.91	77.83	77.54	76.48
Naïve Bayes	80.68	80.64	80.62	80.58	80.68	80.64	80.57	80.45

4.3.1 Emotion Distribution

Using the lexicon-based method, fear, anger, and trust emerged as the most dominant emotions expressed in the tweets. The emotion distribution, shown in Figure 2, highlights widespread anxiety and concern surrounding the conflict, particularly relating to safety, security, and institutional trust. Less prominent emotions such as joy, surprise, and disgust appeared only intermittently, suggesting limited optimism or emotional outrage in the broader conversation.

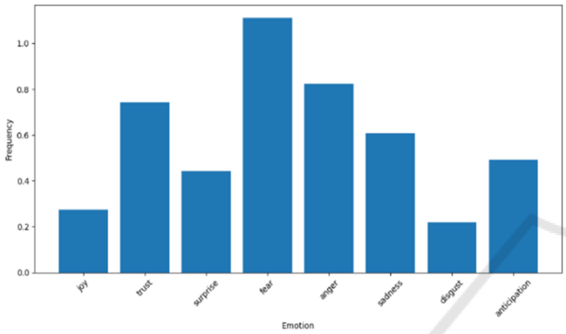

Figure 2: Overall emotion distribution in tweets.

4.3.2 Emotion Word Cloud

To further illustrate the language associated with emotional expression, an emotion word cloud was generated and presented in Figure 3.

This visualization demonstrated the frequency and emotional tone of words associated with each of the eight core emotions. Terms such as "invade," "war," "attack," and "freedom" were particularly prominent under fear and anger, whereas words like "hope" and "peace" featured in association with trust and anticipation, revealing the duality of concern and resilience in user narratives.

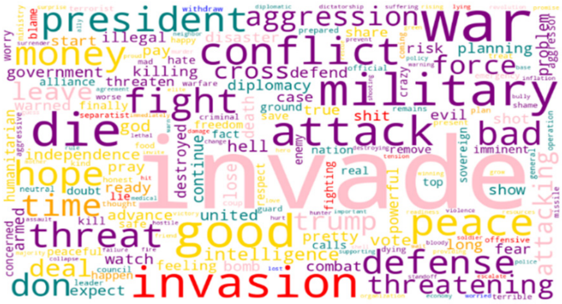

Figure 3: Emotion word cloud.

4.3.3 Top Words per Emotion

Table 4 presents the most frequently occurring words per emotional category. A notable overlap was observed between terms linked to fear, anger, and sadness, suggesting that these emotions often co-occurred in the same discursive contexts. Additionally, the presence of future-oriented words under anticipation (e.g., "plan," "ready") and references to leadership and unity under trust (e.g., "president," "nation") reflected a more complex spectrum of emotional reactions beyond simple negativity.

Table 4: Most frequent emotion-linked words.

Emotion	Top Words
Joy	good, peace, hope, money, intelligence, true, deal, freedom, love, pretty
Trust	president, done, good, peace, united, show, ground, nation, real, hope
Surprise	invade, trump, good, leave, hope, money, deal, attacking, guess, warned
Fear	invade, war, military, die, attack, fight, threat, conflict, aggression, force, defence, government
Anger	Invade, Invasion, attack, fight. Threat, conflict, aggression, force, defence, fighting
Sadness	Invade, die, conflict, leave, bad, problem, cross, withdraw, case, kill
Disgust	Bad, threatening, shit, attacking, hell, blame, separation, loss, enemy, illegal
Anticipation	Time, good, peace, start, long, defence, ready, happen, hope, plan

4.3.4 Comparative Evaluation of Emotion Detection Models

To evaluate the predictive ability of the models for emotion detection, we benchmarked the performance of the lexicon-based and transformer-based models on a manually labelled sample of 300 randomly selected tweets. Due to the resource-intensive nature of human annotation for over a million tweets, this subsample served as a reliable benchmark set. We performed 5-fold cross-validation to compare the effectiveness of both models based on Hamming Loss (HL), Accuracy (1 − HL), Precision, Recall, and F1 Score (macro-averaged). Table 5 shows the fold-wise evaluation results for emotion detection models.

Table 5: Fold-wise evaluation results.

Fold	Model	HL	Acc	Precision	Recall	F1
1	Transformer	0.09	0.90	0.44	0.40	0.36
	Lexicon	0.17	0.82	0.21	0.37	0.25
2	Transformer	0.09	0.90	0.48	0.37	0.37
	Lexicon	0.18	0.81	0.19	0.36	0.24
3	Transformer	0.08	0.91	0.40	0.44	0.39
	Lexicon	0.21	0.78	0.16	0.31	0.17
4	Transformer	0.07	0.92	0.45	0.45	0.44
	Lexicon	0.21	0.78	0.16	0.38	0.20
5	Transformer	0.07	0.92	0.40	0.47	0.40
	Lexicon	0.15	0.84	0.20	0.36	0.26

Additionally, Table 6 presents the average evaluation scores across the five folds for both the Transformer-based and Lexicon-based emotion classification models.

Table 6: Average evaluation scores across the five folds.

Model	HL	Accu	Precision	Recall	F1
Transformer	0.08	0.91	0.48	0.44	0.43
Lexicon	0.18	0.81	0.19	0.36	0.23

The transformer-based model significantly outperformed the lexicon-based model across all evaluation metrics. It achieved an average accuracy of 91% with a notably lower Hamming Loss of 0.08, compared to 81% accuracy and a Hamming Loss of 0.18 for the lexicon-based approach. Additionally, the macro-averaged F1 score for the transformer model was nearly double that of the lexicon model, indicating its superior ability to balance precision and recall across the full range of emotional categories.

To further investigate the classification behavior of each model, confusion matrices were generated and aggregated across the five folds. Figures 4 and 5 present these matrices for the transformer-based and lexicon-based models, respectively.

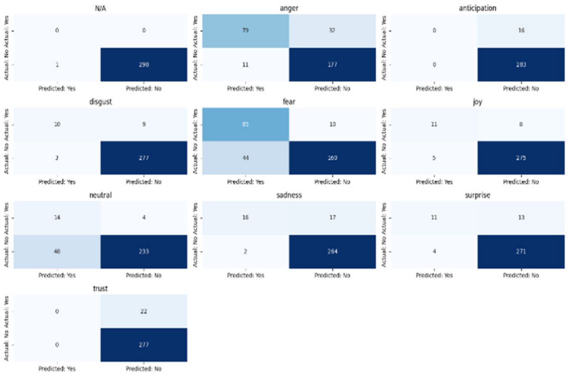

Figure 4: Transformer-based model – per-label confusion matrices.

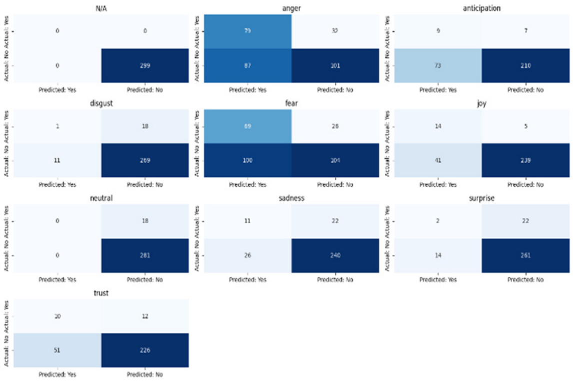

Figure 5: Lexicon-based model – per-label confusion matrices.

Moreover, Table 7 compares the performance of the emotion prediction models based on the True Positives, False Positives, False Negatives, and True Negatives for each emotion.

Table 7: True Positive, False Positive, False Negative, and True Negative values for each emotion.

Emotion	Model	TP	FP	FN	TN
Anger	Transformer	79	11	32	177
	Lexicon	79	87	32	101
Fear	Transformer	85	44	10	160
	Lexicon	69	100	26	104
Joy	Transformer	11	5	8	275
	Lexicon	14	41	5	239
Sadness	Transformer	16	2	17	264
	Lexicon	11	26	22	240
Disgust	Transformer	10	3	19	277
	Lexicon	1	11	18	269
Surprise	Transformer	11	4	13	271
	Lexicon	2	14	22	261
Trust	Transformer	0	0	22	277
	Lexicon	10	51	12	226
Anticipation	Transformer	0	0	16	283
	Lexicon	9	73	7	210
Neutral	Transformer	14	48	4	233
	Lexicon	0	0	18	281

The transformer model was especially effective in identifying fear and anger, with True Positive values of 85 and 79, respectively. It also showed moderate success in detecting emotions like joy, sadness, and surprise, with relatively low false positive counts, indicating reliable contextual interpretation. However, it struggled to identify more subtle emotional expressions such as trust and anticipation, both of which had zero True Positives and high False Negative rates.

In contrast, the lexicon-based model frequently misclassified emotions, especially anger and fear, resulting in high False Positive rates. Its performance on emotions like joy, trust, and disgust was also notably poor, with misclassifications stemming from the lack of contextual understanding. Overall, the confusion matrix results reinforce the quantitative findings and suggest that transformer-based emotion detection models are better equipped to capture the complexity of emotional discourse on social media.

While the transformer model performed well for emotions like fear and anger, it struggled with subtler categories such as trust and anticipation, likely due to data sparsity and their abstract, context-dependent nature. Future work could fine-tune models on larger, more balanced datasets and apply contextual augmentation techniques like paraphrasing or prompt-based learning to improve detection of subtle emotions.

5 DISCUSSIONS

This study sought to analyse public sentiment and emotion in response to the Russia-Ukraine conflict by leveraging a large-scale Twitter dataset and applying both machine learning and transformer-based techniques. The results demonstrate clear differences not only in public attitudes but also in the effectiveness of various computational approaches to interpreting online discourse.

The sentiment classification revealed a marked skew toward negativity in early 2022, with negative tweets outnumbering positive ones by a significant margin. This finding aligns with the tense geopolitical context at the time, marked by uncertainty, violence, and widespread concern. The high proportion of neutral tweets suggests that many users chose to report or retweet news without adding personal commentary, or expressed sentiments that did not fall neatly into traditional polarity categories.

For sentiment prediction, the Feedforward Deep Learning model paired with Count Vectorizer showed the strongest performance, capturing nuanced sentiment better than traditional models and TF-IDF features. These results challenge the conventional reliance on Naïve Bayes and Support Vector Machines, which underperformed in this context.

In emotion classification, fear and anger were the most common emotions, reflecting the conflict's psychological impact. Trust and anticipation appeared less often but suggested resilience and hope. The overlap between fear, anger, and sadness terms highlights the emotionally complex nature of public reactions to conflict.

In terms of model performance for emotion detection models, the transformer-based model outperformed the lexicon-based approach across all emotion metrics, with confusion matrix analysis highlighting its superior contextual sensitivity and lower false positives. However, it still struggled with low-frequency or abstract emotions like trust and anticipation, suggesting a need for further model refinement or data augmentation.

Moreover, the superior performance of transformer models over lexicon-based approaches stems from their ability to capture rich contextual embeddings. Unlike lexicon methods that analyse isolated words, transformers like DistilRoBERTa process entire sequences, enabling them to detect nuances such as sarcasm, irony, negation, and polysemy—common in social media. By modelling word interdependencies, transformers offer a more accurate understanding of emotional tone, making them better suited for informal and complex communication.

5.1 Limitations

The limitations of this study are important to acknowledge. The dataset was restricted to English-language tweets, which potentially omits culturally specific expressions of sentiment and emotion that are communicated in other languages. This language constraint may lead to an under-representation of viewpoints from non-English-speaking users, particularly those directly affected by the Russia-Ukraine conflict in Eastern Europe. As a result, the findings may disproportionately reflect the perspectives of English-speaking users, which could skew interpretations of global sentiment. To improve generalizability, future research should explore the integration of multilingual sentiment and emotion analysis using models such as multilingual BERT (mBERT) or XLM-Roberta. These models are designed to handle a wide array of languages and could help capture a more comprehensive and culturally diverse picture of public discourse during international crises.

Furthermore, while 300 manually annotated tweets were used for validation, a larger benchmark set could improve generalizability. Additionally, focusing solely on Twitter excludes sentiments from other platforms like Facebook, Telegram, and Reddit, which may reflect different demographics and viewpoints.

5.2 Knowledge Contributions

This study advances sentiment and emotion analysis, NLP, and computational social science by comparing traditional machine learning with deep learning models, highlighting the superior performance of transformers on large-scale social media data. It also compares Count Vectorizer and TF-IDF for handling informal, short-form texts like tweets.

By combining sentiment classification with multi-label emotion detection, this study offers a more nuanced understanding of public reactions during a global crisis. It goes beyond simple sentiment categories to reveal complex emotions like fear, anger, trust, and hope, capturing the psychological and affective dimensions of conflict-related discourse.

Crucially the study provides strong evidence that transformer-based models outperform traditional lexicon methods in detecting subtle, overlapping, and context-dependent emotions, highlighting the value of using context-aware models for a more complete analysis of public discourse.

Another contribution is the creation of a large-scale dataset of over 1.4 million tweets from the early stages of the Russia-Ukraine war, which supports the study's analysis and offers a valuable resource for future research on public sentiment during international crises.

Beyond its technical contributions, this research provides actionable insights for policymakers, media analysts, and humanitarian organizations, helping them better understand public opinion and emotions—key for shaping communication and policy during global crises.

6 CONCLUSION

This study analysed public sentiment and emotions toward the Russia-Ukraine conflict using a large set of English-language tweets. Sentiment classification with traditional and modern models showed that the deep learning model with Count Vectorizer achieved the highest accuracy (89.58%), demonstrating the strength of advanced models in capturing large-scale, real-time social media sentiment.

To explore deeper emotional nuances, the study compared a lexicon-based model (NRC) and a transformer-based model (DistilRoBERTa). Benchmark results showed the transformer model significantly outperformed the lexicon approach across all metrics, highlighting its strength in multilabel classification and detecting nuanced, context-dependent emotions.

By combining sentiment and emotion analysis, this research presents a more holistic view of online public discourse during geopolitical crises. It provides valuable insights not only for the advancement of NLP techniques but also for stakeholders—such as policymakers, media analysts, and humanitarian agencies—seeking to understand and respond to public opinion during times of conflict.

REFERENCES

Ahmad, M., Aftab, S., Muhammad, S.S. and Ahmad, S. (2017). Machine learning techniques for sentiment analysis: a review. *Int. J. Multidiscip. Sci. Eng*, 8(3), p.27.

Ahuja, S. and Dubey, G. (2017). Clustering and sentiment analysis on Twitter data. In *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)* (pp. 1–5).

Al Maruf, A. et al. (2022). Emotion Detection from Text and Sentiment Analysis of Ukraine Russia War using Machine Learning Technique. *International Journal of Advanced Computer Science and Applications*, 13(12).

Alhindi, T., Ghosh, S., & Greene, D. (2018). Emotion Detection in English and Arabic Tweets Using Deep Learning. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*.

Amirhosseini, M.H., Kazemian, H.B., Ouazzane, K., and Chandler, C. (2018). Natural Language Processing approach to NLP Meta model automation. *2018 International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, Brazil, pp. 1–8. https://doi.org/10.1109/IJCNN.2018.8489609.

Asghar, M.Z., Khan, A., Bibi, A., Kundi, F.M. and Ahmad, H. (2017). Sentence-level emotion detection framework using rule-based classification. *Cognitive Computation*, 9(6), pp.868–894.

Aung, K.Z. and Myo, N.N. (2017). Sentiment analysis of students' comment using lexicon-based approach. In *2017 IEEE/ACIS 16th International Conf. on Computer and Information Science (ICIS)* (pp. 149–154).

Baid, P., Gupta, A. and Chaplot, N. (2017). Sentiment analysis of movie reviews using machine learning techniques. *International Journal of Computer Applications*, 179(7), pp.45–49.

Baker, M.R., Taher, Y.N., and Jihad, K. (2023). Prediction of People Sentiments on Twitter Using Machine Learning Classifiers During Russian Aggression in Ukraine. *International Journal of Computers and Information Technology*, 9(3), pp.183–204.

Berka, P. (2020). Sentiment analysis using rule-based and case-based reasoning. *Journal of Intelligent Information Systems*, 55(1), pp.51–66.

Bhatt, S., Ghazanfar, M., and Amirhosseini, M. (2023). Sentiment-Driven Cryptocurrency Price Prediction: A Machine Learning Approach Utilizing Historical Data and Social Media Sentiment Analysis. *Machine Learning and Applications: An International Journal (MLAIJ)*, 10(2/3), 1–15.

Bhavitha, B.K., Rodrigues, A.P. and Chiplunkar, N.N. (2017). Comparative study of machine learning techniques in sentimental analysis. In *2017 International Conf. on Inventive Communication and Computational Technologies*, pp. 216–221.

Bonta, V. and Janardhan, N.K.N. (2019). A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology*, 8(S2), pp.1–6.

Chaturvedi, I., Cambria, E., Welsch, R.E. and Herrera, F. (2018). Distinguishing between facts and opinions for sentiment analysis: survey and challenges. *Information Fusion*, 44, pp.65–77.

Chekima, K., Alfred, R. and Chin, K.O. (2017). Rule-based model for Malay text sentiment analysis. In *International Conference on Computational Science and Technology* (pp. 172–185).

Dang, N.C., Moreno-García, M.N. and De la Prieta, F. (2020). Sentiment analysis based on deep learning: a comparative study. *Electronics*, 9(3), p.483.

Dehghani, M., Johnson, K.M., Garten, J., Boghrati, R., Hoover, J., Balasubramanian, V., Singh, A., Shankar, Y., Pulickal, L., Rajkumar, A. and Parmar, N.J. (2017). TACIT: an open-source text analysis, crawling, and interpretation tool. *Behavior Research Methods*, 49(2), pp.538–547.

Dwivedi, R.K., Aggarwal, M., Keshari, S.K. and Kumar, A. (2019). Sentiment analysis and feature extraction using rule-based model (RBM). In *International Conf. on Innovative Computing and Communications* (pp. 57–63).

Gupta, A., Joshi, R., & Jain, S. (2022). Emotion detection in text: A review. *ACM Computing Surveys*, 55(1), 1–38.

Hartmann, J. (2022). *j-hartmann/emotion-english-distilroberta-base* [Model]. HuggingFace. Available at: https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

Hasan, A., Moin, S., Karim, A. and Shamshirband, S. (2018). Machine learning-based sentiment analysis for Twitter accounts. *Mathematical and Computational Applications*, 23(1), p.11.

Huq, M.R., Ahmad, A. and Rahman, A. (2017). Sentiment analysis on Twitter data using KNN and SVM. *International Journal of Advanced Computer Science and Applications*, 8(6).

Hussein, D.M.E.D.M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University - Engineering Sciences*, 30(4), pp.330–338.

Jagdale, R.S., Shirsat, V.S. and Deshmukh, S.N. (2019). Sentiment analysis on product reviews using machine learning techniques. In *Cognitive Informatics and Soft Computing* (pp. 639–647).

Khoo, C.S. and Johnkhan, S.B. (2018). Lexicon-based sentiment analysis: comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44(4), pp.491–511.

Liao, S., Wang, J., Yu, R., Sato, K. and Cheng, Z. (2017). CNN for situations understanding based on sentiment analysis of Twitter data. *Procedia Computer Science*, 111, pp.376–381.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint* arXiv:1907.11692.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3), 436–465.

Mohammad, S.M. (2017). Challenges in sentiment analysis. In *A Practical Guide to Sentiment Analysis* (pp. 61–83). Springer, Cham.

Mostafa, M.M. (2019). Clustering halal food consumers: a Twitter sentiment analysis. *International Journal of Market Research*, 61(3), pp.320–337.

Ray, P. and Chakrabarti, A. (2020). A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis. *Applied Computing and Informatics*.

Sailunaz, K., & Alhajj, R. (2019). Emotion and sentiment analysis from Twitter text. *Journal of Computational Science*, 36, 101003.

Stine, R.A. (2019). Sentiment analysis. *Annual Review of Statistics and Its Application*, 6, pp.287–308.

Vashishtha, S. and Susan, S. (2019). Fuzzy rule based unsupervised sentiment analysis from social media posts. *Expert Systems with Applications*, 138, p.112834.

Wang, Y., Kim, K., Lee, B. and Youn, H.Y. (2018). Word clustering based on POS feature for efficient Twitter sentiment analysis. *Human-Centric Computing and Information Sciences*, 8(1), pp.1–25.

Yadollahi, A., Shahraki, A. G., & Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2), 1–33.

Zhang, L., Wang, S. and Liu, B. (2018). Deep learning for sentiment analysis: a survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), p.e1253.

Zimmer, M. and Proferes, N.J. (2014). A topology of Twitter research: disciplines, methods, and ethics.