JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

"To Stand with Ukraine is to stand with Humanity": Sentiment Analysis using Machine Learning with NLP

[1]**Rayees Ahmad**

[1]Research Scholar,
IIPS, DAVV, Takshshila Campus,
Indore (M. P.), India

[2]**Yasmin Shaikh**

[2]Assistant Professor,
IIPS, DAVV, Takshshila Campus,
Indore (M. P.), India

[3]**Sanjay Tanwani**

[3]Professor & Head,
School of CS & IT, DAVV,
Takshshila Campus, Indore (M. P.), India

Abstract

Social media platforms and Microblogging sites can be used to gather public opinion and sentiment on a range of topics, including the current state of affairs in war-torn countries. During a crisis, Online Social Networks (OSNs) play a critical role in information sharing. The information gathered during such a crisis, public opinion and sentiments on a large scale can be reflected. Twitter, in particular, contains a significant quantity of geo tagged tweets, allowing for sentiment analysis over time and geography.

The primary goal of this research study is to harness the power of social media to monitor, examine, and analyze public opinion on a recent "Russia's Invasion on Ukraine", as public opinion is crucial in forming government policy. By delving deeper into social media, one may readily study people's behavior on a variety of subjects and policies, which would be impossible to do otherwise using traditional sources.

In this research paper, it is aimed to classify the viewpoint as Positive, Negative, or Neutral by using Machine learning techniques (Lexicon based) with Natural Language Processing (NLP). The findings of this study can assist various organizations and stockholders in improving their political strategies and commercial decision-making for current and future intents by utilizing social media networks as a valuable source of knowledge.

Keywords:

Sentiment Analysis, Natural Language Processing, Machine Learning, Twitter, Ukraine, Russia, Political reviews

1.INTRODUCTION

Technology improvements have a significant impact on every aspect of life, including politics. Artificial intelligence has showed potential benefits in politics by analyzing data and making decisions. People can communicate their opinion, attraction, and feelings through social media, which is a simple technique of communication technology.

For the people, online social networks have become a popular communication tool. On social media, people are continually expressing their views. As a result, social networking sites provide a rich supply of information for opinion mining. We can tell how well a product is doing in the present market by studying the numerous opinions offered on such sites. In this research, we propose a system that mines user views on products or services using the popular microblogging website Twitter. Our system proposes a method for extracting data from Twitter and analyzing it linguistically.

The purpose of the paper is to extract a variety of emotions. Twitter has grown in popularity as a platform for people to share their thoughts and ideas on a variety of topics. Twitter social media analysis can be used to determine which sentiments are prevalent. This information will be used to make a strategic decision. It also helps to categorize people's feelings and affections. Whether it's apparent, conflicting, or neutral, the information was pre-processed. With the use of noise cancellation, the noise can be removed. A number of sentiment analysis techniques are applied on the available to data to get the desired results.

In this research, twitter API as the platform for analyzing citizen reactions to events in Ukraine. Twitter API was used for text mining to reveal political sentiments collected based on particular keywords from the Twitter platform between February 24 and May 8 2022. Sentiment analysis correlated to study the opinion mining, a process of identification, classification and result generation of individual thoughts about any product or event over the internet either is positive, negative or neutral Liu et al. [1].

The findings of this study can assist palpitations, firms, and stockholders in using social media as a valuable source of information for better political strategies and business decision-making in the present and future. It presents a practical strategy as well as a case study as an example to help researchers implement sentiment analysis techniques more effectively.

2. BACKGROUND AND RELATED WORK

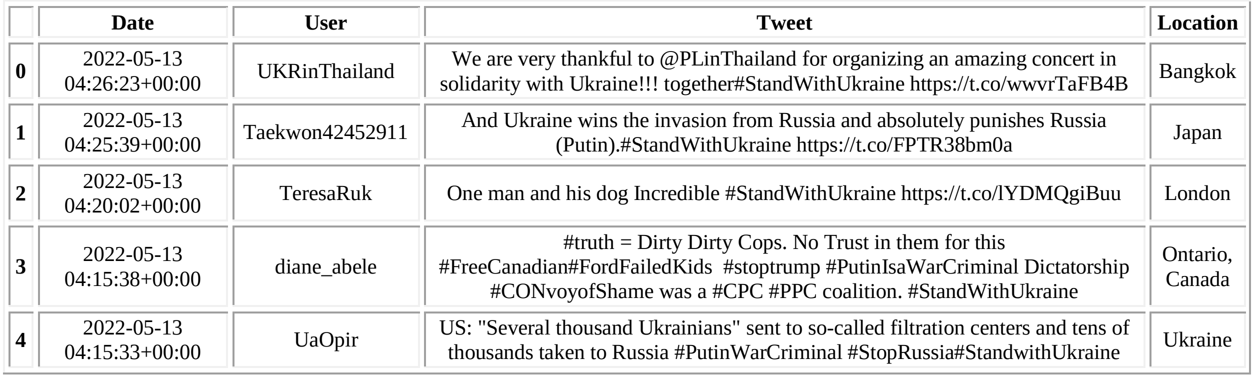
As users continue to post a large amount of textual information on various social media sites, there is a growing interest in using automatic methods such as text mining and sentiment analysis to process large amounts of user-generated data and extract meaningful knowledge and insights [2]. As an emerging technology, text mining aims to extract meaningful information from a large number of textual documents quickly [3]. Text mining is focused on finding useful models, trends, patterns, or rules from unstructured textual data [4]. Sentiment analysis is an emerging research field of NLP for the extraction of people's opinions, thoughts, and views [9]. It is highlighting on text classification and relates to text mining. Detection of sentiment over the web and social media is difficult because of the bulk amount of data [5]. With advancements in social networking after the year 2000, the Internet's users have increasingly been posting their subjective thoughts and feeling on the social platforms. People's connections over the internet produced a bulk amount of data about their opinions and issues. Sentiment analysis and text mining deal with unstructured data [6]. Sentiment analysis involves classifying opinion in the text of positive, negative and neutral. Sentiment analysis is a method of finding out whether a piece of text is positive, negative or neutral [7]. This technique is also called opinion mining, by deriving the attitude or opinion of data [8]. Sentiment analysis can be categorized into the following levels, document level [9], Sentence level [10], and Tweet-level [11]. Lexicon method is a principal technique of sentiment, which is unsupervised and deals with text classification into predefined sentiment classes. Lexicon is base for the calculation of sentiment scores and polarity of text data [12]. But the lexicon method can't provide high performance due to the terms in text data, which can be resolved by creating a context-specific lexicon. Lexicon-based provide an opportunity to make datasets well concerning a dictionary. Currently, a microblogging site, Twitter is most popular for research in various fields this high popularity of Twitter can be used for many purposes like political campaigns, learning tool, and for emergency too. For instance, Pak and Paroubek [14] proposed a model to classify the tweets as an object, positive and negative. In this model they created a Twitter corpus by getting tweets using Twitter API, they developed a sentiment classifier based on Naïve Bayes method which uses POS-tags and N-gram. Similarly, Kim inspected social media user's response after the 2016 flood in Louisiana and observed that during an emergency the information about disaster diffused over social networks [19].

3. RESEARCH METHOD

Russia's invasion on Ukraine is taken as the subject of the case study. Using the Python programming language, analysis is carried out on one of the most well-known social networking site, Twitter. The primary data is gathered via Twitter API. Opinion lexicon techniques are used to analyze the tweets. The sentiments of tweets are classified using a Lexicon-based technique.

a. Data set Description

The dataset for this study's experiments is gathered from Twitter. The Tweepy library and a developer account are used to extract the tweets for this purpose [22]. Several different keywords and hashtags such as '#standforRussia' '#standwithRussia', '#RussiaUkraineCrisis' etc. are used to search the relevant tweets. The collected tweets are posted by Twitter users from 24 February 2022 to May 8 2022. For sentiment analysis, a total of 110,000 tweets are retrieved, with a few sample tweets listed in Table 1.The tweet information is stored in CSV files along with the user who posted the tweet, Date at what time it was posted and The geo-location of where the tweets were posted is also captured. The extracted dataset's location distribution of Twitter users is shown in Figure 1.



	Date	User	Tweet	Location
0	2022-05-13 04:26:23+00:00	UKRinThailand	We are very thankful to @PLinThailand for organizing an amazing concert in solidarity with Ukraine!!! together#StandWithUkraine https://t.co/wwvrTaFB4B	Bangkok
1	2022-05-13 04:25:39+00:00	Taekwon42452911	And Ukraine wins the invasion from Russia and absolutely punishes Russia (Putin).#StandWithUkraine https://t.co/FPTR38bm0a	Japan
2	2022-05-13 04:20:02+00:00	TeresaRuk	One man and his dog Incredible #StandWithUkraine https://t.co/lYDMQgiBuu	London
3	2022-05-13 04:15:38+00:00	diane_abele	#truth = Dirty Dirty Cops. No Trust in them for this #FreeCanadian#FordFailedKids #stoptrump #PutinIsaWarCriminal Dictatorship #CONvoyofShame was a #CPC #PPC coalition. #StandWithUkraine	Ontario, Canada
4	2022-05-13 04:15:33+00:00	UaOpir	US: "Several thousand Ukrainians" sent to so-called filtration centers and tens of thousands taken to Russia #PutinWarCriminal #StopRussia#StandwithUkraine	Ukraine

Figure 1. Extracted datasets location distribution

The extracted tweets contain various words that appear in the majority of the tweets, such as Russia, Ukraine, and crisis.

b. Proposed Methodology

This research examines the attitude of tweets about Russia's invasion on Ukraine. Experiments are carried out on a Windows 10 PC with an Intel Corei7 7th Generation processor. For implementing Machine learning and deep learning models, Jupyter Notebook, NLP and Python language are used. Experiments are conducted using the Scikit library, Keras, TensorFlow, Genism, Textblob, and NLTK libraries. Figure 3 depicts the architecture of sentiment Analysis techniques.

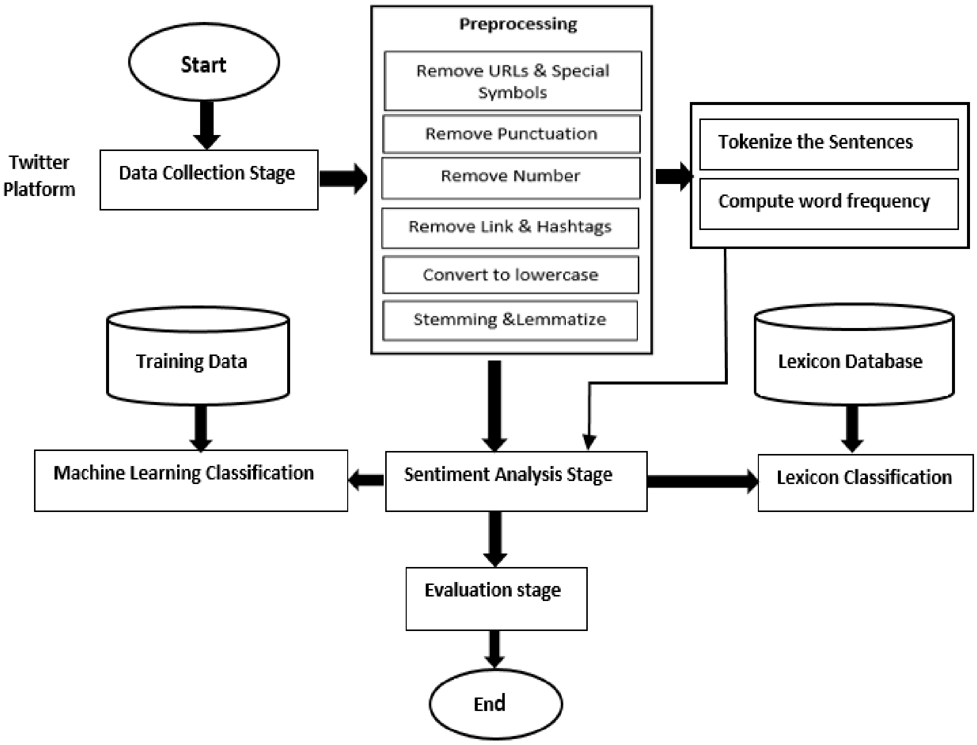
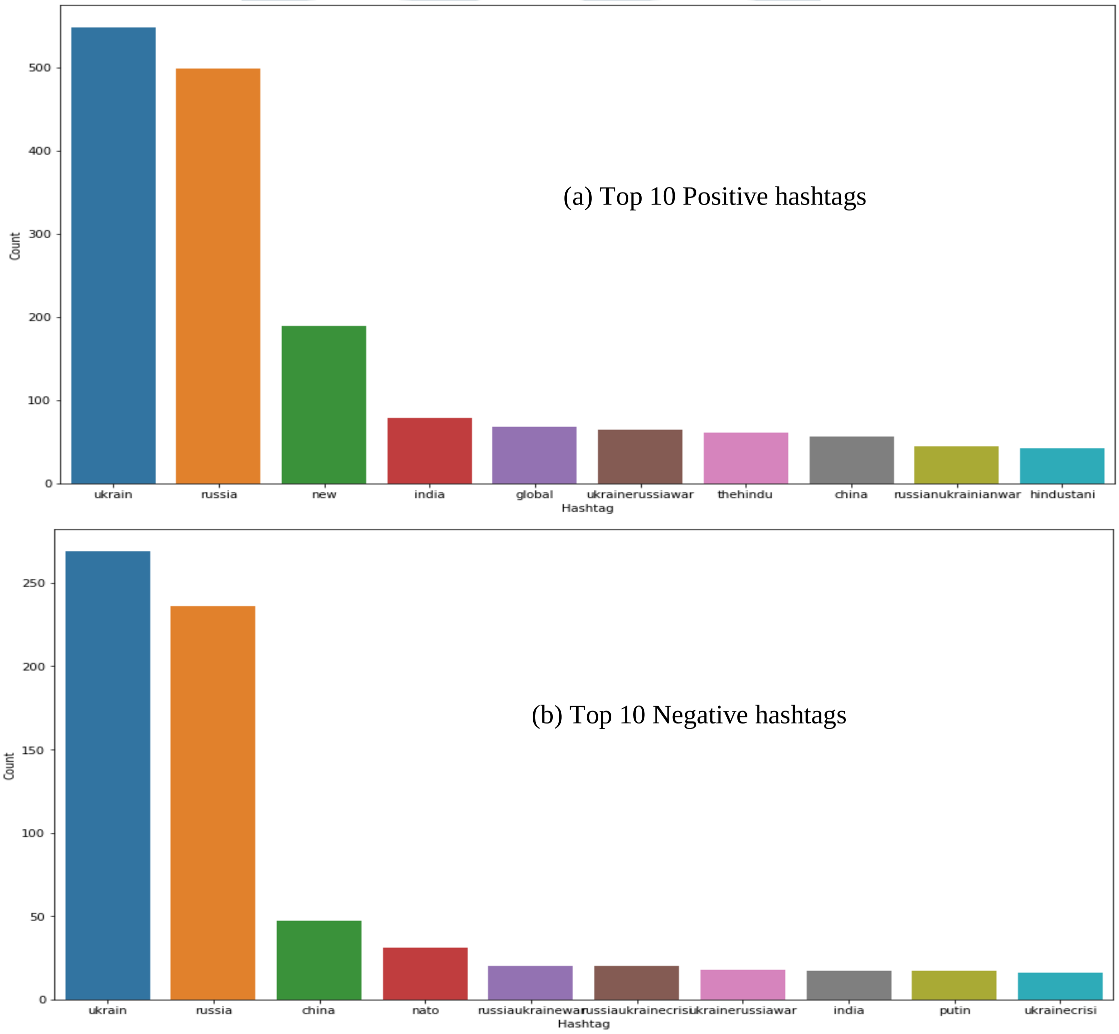


Figure 2. Architecture of Sentiment Analysis techniques

Following data extraction, the first step is to remove any unwanted or excessive information from the data that does not contribute to the target class prediction. A preprocessing pipeline is used to clean the acquired data for this purpose [25], [26]. The steps that were followed in the sentiment analysis are shown in figure 2 above.

- **Remove URLs and Special Symbols:** Uniform Resource Locators (URL's) in a text are references to a location on the web, but do not provide any additional information. Special characters are non-alphanumeric characters which are most often found in comments, references, currency numbers etc. These characters add no value to text-understanding and cause algorithmic noise. As a result, URLs and Special symbols are removed to reduce feature space complexity.
- **Punctuation and number removal:** Punctuation is an important aspect of a sentence since it helps human readers understand it. Punctuation, on the other hand, adds complexity to machine learning models and increases the size of the feature vector. Punctuation is eliminated from tweets due to the fact that it does not contribute to the training process.
- **Hashtags and link removal:** To decrease the dataset's complexity, tweets with tags and links that aren't beneficial for training the models are eliminated. Regular expressions are used to remove tags and links.



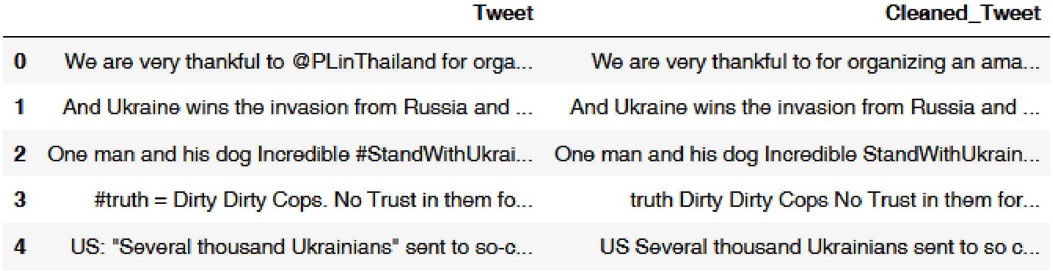
(a) Top 10 Positive hashtags

(b) Top 10 Negative hashtags

Figure 3. Positive & Negative hashtags

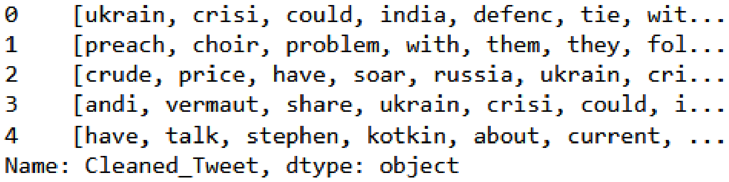
- **Convert to Lowercase:** Because machine learning models are case sensitive, the difference in case causes difficulty in model training because the words 'Peace,' and 'PEACE' are all deemed to be different terms. It increases both the size of the feature space and the processing time. Each character is converted to its lower case using a Python built-in function called conversion to lower case.

- **Remove Stop words:** Stop words are a vital aspect of a sentence's readability and meaning. These are brief, meaningless terms like the, in, he, it, an, and others that don't help train machine learning models. As a result, stop words are deleted to reduce feature space complexity.



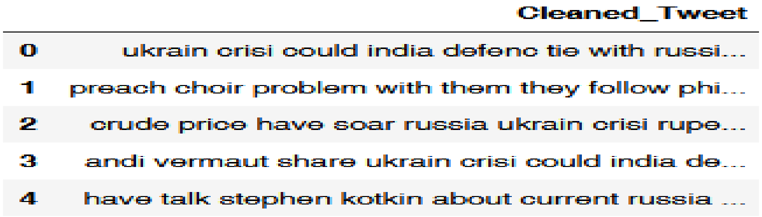
	Tweet	Cleaned_Tweet
0	We are very thankful to @PLinThailand for orga...	We are very thankful to for organizing an ama...
1	And Ukraine wins the invasion from Russia and ...	And Ukraine wins the invasion from Russia and ...
2	One man and his dog Incredible #StandWithUkrai...	One man and his dog Incredible StandWithUkrain...
3	#truth = Dirty Dirty Cops. No Trust in them fo...	truth Dirty Dirty Cops No Trust in them for...
4	US: "Several thousand Ukrainians" sent to so-c...	US Several thousand Ukrainians sent to so c...

- **Stemming and lemmatization:** Stemming reduces words to their root forms, such as 'goes,' 'going,' and 'gone,' which are all variations of the word 'go.' Machine learning models interpret these as different words if not handled appropriately, therefore stemming is used to convert them all to their fundamental form 'go'. Although stemming and lemmatization are similar, lemmatization is generally more successful. Stemming gets rid of a few characters, which can lead to mistakes.



```
0    [ukrain, crisi, could, india, defenc, tie, wit...
1    [preach, choir, problem, with, them, they, fol...
2    [crude, price, have, soar, russia, ukrain, cri...
3    [andi, vermaut, share, ukrain, crisi, could, i...
4    [have, talk, stephen, kotkin, about, current, ...
Name: Cleaned_Tweet, dtype: object
```

The results of preprocessing processes performed on sample tweets from the gathered dataset is shown in figure below:



	Cleaned_Tweet
0	ukrain crisi could india defenc tie with russi...
1	preach choir problem with them they follow phi...
2	crude price have soar russia ukrain crisi rupe...
3	andi vermaut share ukrain crisi could india de...
4	have talk stephen kotkin about current russia ...

Figure 4. Results of Using Preprocessing steps

4. EXPERIMENTATION & RESULTS

A total of 110,000 tweets were retrieved for this study, but only 59927 tweets, or 54% of the total, are included in the sentiment analysis after data cleaning. Many duplicate tweets and retweets were found in the retrieved data, and they are removed from the dataset. "Retweets" and duplicates do not present fresh opinions, but they would alter the overall results.

Figure 5 depicts the Word Cloud for the entire dataset, which shows the most frequently used terms in the extracted tweets about Russia’s invasion on Ukraine.

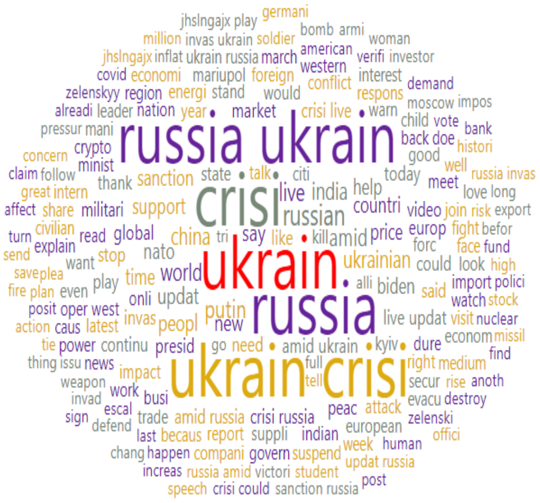


Figure 5. Word cloud of the collected

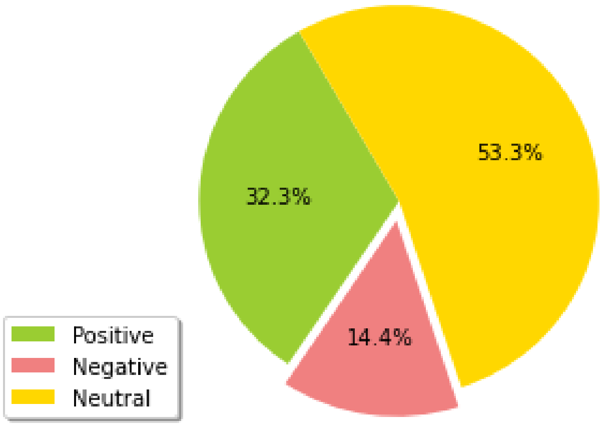


Figure 6. Sentiment Distribution

Positive comments are usually always higher than negative comments on Twitter between 15 March 2022 and 30 March 2022, however from 13 April 2022 to 25 April 2022, Neutral comments are higher than positive remarks. The cause for this was a brief period of interim peace between the Russia and the Ukrainian group.

The biggest number of tweets originate from US, UK, Ukraine, Russia, India and China according to volume analysis using geo-tagged tweets. Positive tweets come from first world countries Like US, UK, Canada and France, negative comments on the other hand come from second world countries like Russia, China etc. Non Aligned countries like India, Pakistan and Saudi Arabia remain neutral.

The sentiment analysis revealed that out of the total amount of tweets, 14.4% were categorized as unfavorable, 53.3% percent as neutral, and 32.3% percent as positive. The sentiment analysis findings of Tweets are visualized in Figure 6.

In comparison to other sentiments, the neutral tweet category is always exhibited at the top as shown in Bar chart sentiment distribution in Figure 9. The bar chart displays the count values of the sentiments. Figure 10 depicts the emotion distribution of words on May 5, 2022. While another finding from the graph is that the negative tweet category is always at the bottom of the graph by a wide margin from the subsequent category, positive comments on Twitter are higher than negative comments from April 4 to April 20, 2022.

From the collected dataset, Figure 7 and 8 shows the ratio of positive to negative tweets for the top seven counties. US had the highest percentage of positive tweets and the lowest percentage of negative tweets, indicating US support for Ukraine. The fraction of negative tweets, on the other hand, is larger in china, highlighting political interest tendencies. It also reveals that, for the most part, the attitudes expressed in tweets are neutral, with the greatest percentage of 18% seen in tweets from India.

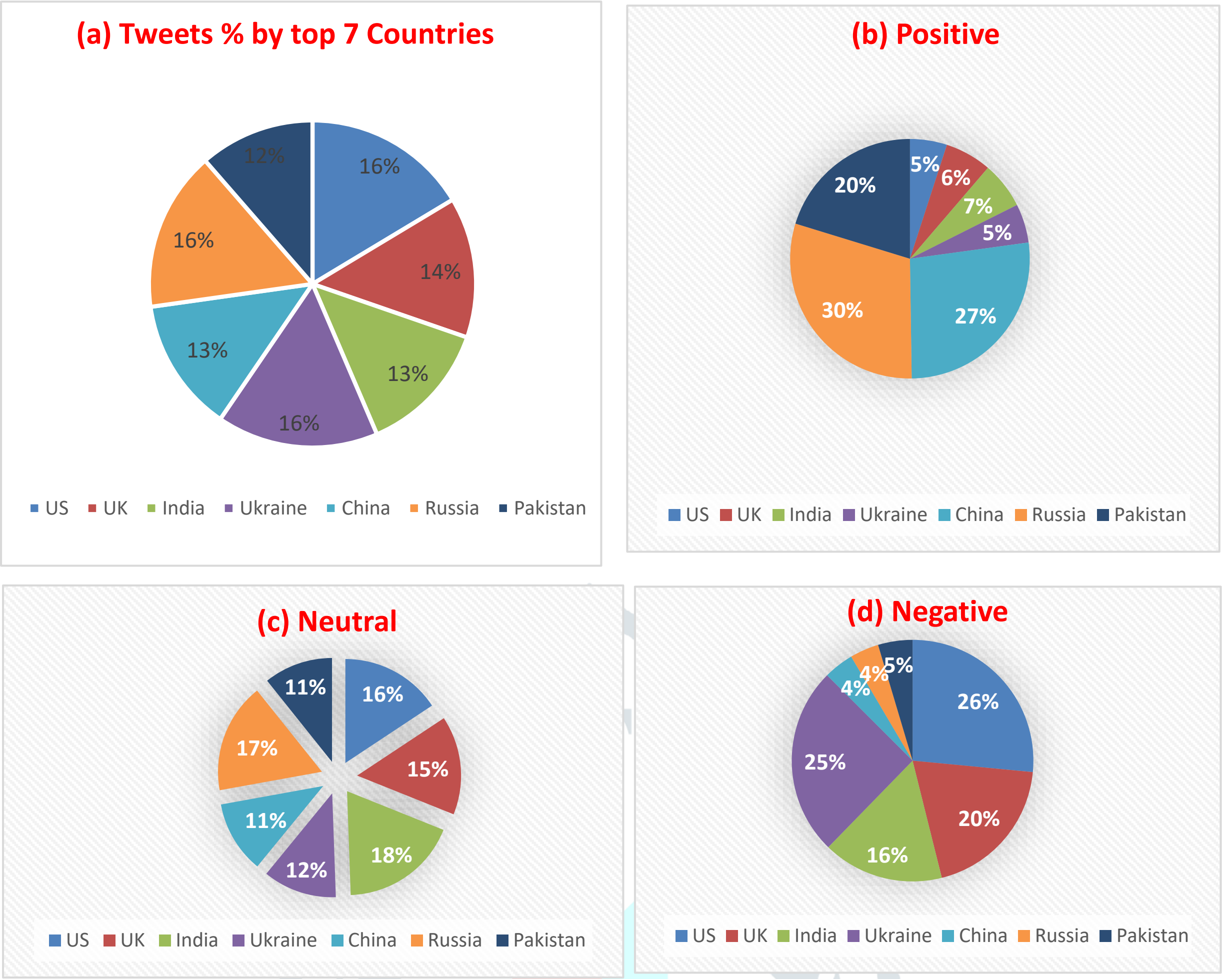


Figure 7. Top seven countries in dataset in terms of number of tweets, a) Tweets Percentage by top seven countries b) Negative tweets ratio c) Neutral tweets ratio d) Positive tweets ratio

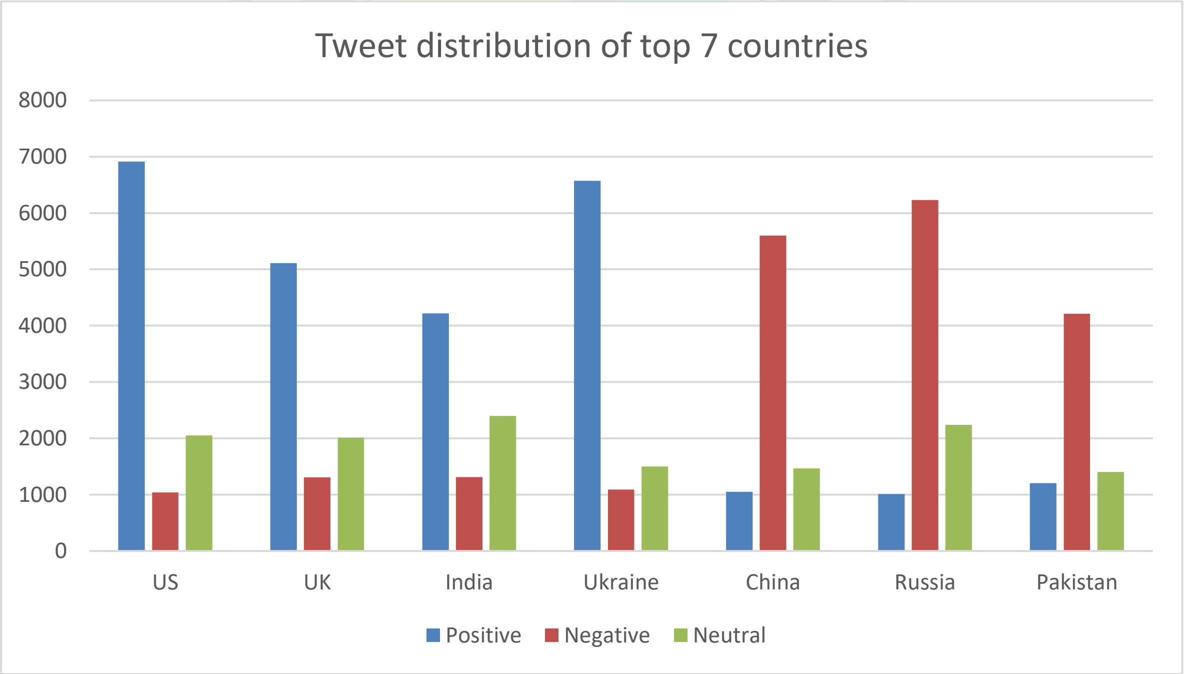


Figure 8. Tweet Distribution for top seven countries with highest number of tweets

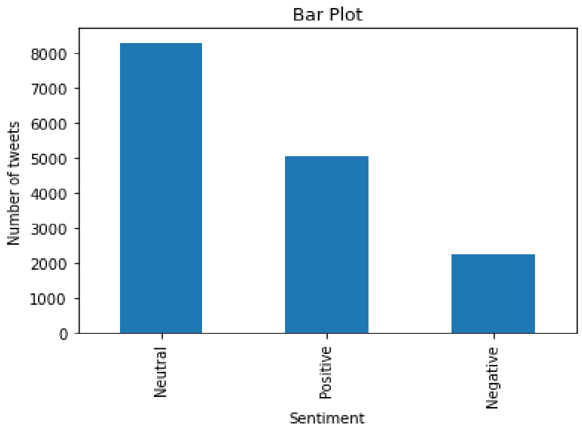


Figure 9. Bar chart displaying count of the sentiments

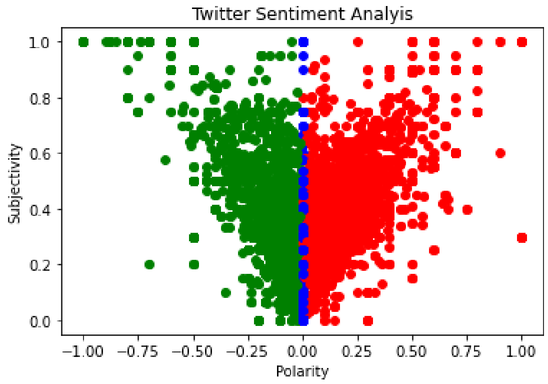


Figure 10. Tweet Sentiment Distribution graph

(a)

Count	Negative	Positive
1	Killed	Winning
2	Destroyed	Trust
3	Defeated	Freedom
4	Slaver	Better
5	Failed	Victory
6	Angry	Conquer
7	Horrifying	Expertise
8	Terror	Easily
9	Pathetic	Free
10	war crime	right

(b)

Country	Positive	Negative
US	Victory, secure, safe, success, love, best,	Worst, painful, terrorism, terror, kill
UK	Power, innocent, charities, right, congratulates	Destroyed, conflict, death, horrifying, Hell, war
India	Peace, loving, success, good, brilliant	Bad, kill, wrong, evil, hypocrisy
Ukraine	Winning, good, safe, best, peace	Worse, terrorist, fears, horrible, defeated
Pakistan	Victory, safe, success, love, best	Bad, kill, wrong, evil, hypocrisy
Russia	Peace, victory, safe, success, love	Defeated, fears, wrong, kill

Figure 11.Mostly used words in tweets impact on negative and positive sentiments, (a) and (b)

The sequence of sentence-based opinions presented in the dataset has been examined to determine which terms are associated with the most unfavorable and positive comments. Figure 11(a) depicts the most frequently used words in tweets with both negative and positive feelings. Additional experiments are carried out on the country level to uncover positive and negative words in addition to the most popular words from the overall twitter datasets. Table 11(b) presents a list of positive and negative adjectives associated with each of the top seven countries with the most tweets.

The sentiment distribution is depicted in Figure 6. Twitter, 5 May 2022. The data suggest that the 8th of May received the most unfavorable comments, with the words "Russia," "killed," "shame," and "assault" being used frequently.

5. CONCLUSION

This study used Twitter data to conduct a series of sentiment analysis on the topic of "Russia's Invasion on Ukraine." And a list of relevant tweets has been collected from the Twitter API. A total of 110,000 tweets were collected for study after a thorough Twitter search. Following data cleaning, 59927 English tweets were utilized in the studies after duplicates, retweets, and tweets with missing information were removed.

According to the results of sentiment analysis, 14.4% of tweets were classed as unfavorable, 53.3 percent as neutral, and 32.3 percent were labelled as positive. Another thing to notice about Figure 4 is that a good tweet category is always at the bottom of the graph, much below the subsequent category.

Individual sentiment scores are used in the overall review to better understand the country's position through time, as well as the most common negative and positive terms to better understand people's difficulties. For example, "peace," "Ukraine," and "secure" are common good terms, whereas "kill," "attack," "Russia," "bomber," and "weapon" are common negative phrases, indicating people's concerns and happiness.

Furthermore, as part of the data cleaning process for this study, Emoji characters were deleted from tweets. Emoji characters are being utilized increasingly regularly in social media messages these days. They've been shown to influence the overall tone of Twitter posts [26].

The study's limitation is that it only looked at English tweets and did not look at other languages' responses. As a result, the results must be confirmed using a huge number of comments written in other languages. For future Research an Urdu sentiment lexicon is intended to establish and compare the results with English tweets, as these two languages will provide us with better results. Also Emoji characters is in our social media is decided to implement in sentiment analysis studies, as adding Emoji characters may help us get more accurate sentiment scores.

The use of Twitter news feeds for analysis was also investigated because they provide crucial temporal information about the situations. This may have boosted the amount of neutral tweets marginally.

REFERENCES

[1]. Liu, B. (2012). "Sentiment analysis and opinion mining.Synthesis Lectures on Human Language Technologies", 5(1), 1–167. doi:10.2200/S00416ED1V01Y201204HLT016.

[2]. Wu He, XinTian, Ran Tao, Weidong Zhang, Gongjun Yan, VasudevaAkula, (2017) "Application of social media analytics: a case of analyzing online hotel reviews", Online Information Review, Vol. 41

[3]. He, W., Zha, S.H. and Li, L. (2013), "Social media competitive analysis and text mining: a case study in the pizza industry," International Journal of Information Management, Vol. 33 No. 3, pp. 464-472.

[4]. He, W., Shen, J., Tian, X., Li, Y., Akula, V., Yan, G. and Tao, R. (2015), "Gaining competitive intelligence from social media data: evidence from two largest retail chains in the world", Industrial Management & Data Systems, Vol. 115 No. 9, pp. 1622-1636.

[5]. Fernadex-Gavilanes, M, Avares-Lopez, T., Juncal-Martinez, J., CostaMontenegro, E., Gonzalez-Castano, F.J., 2016. Unsupervised method online texts. Expert syst Appl. 58, 57-75.

[6]. Oza, K.S., Naik, P.G., 2016. Prediction of online lectures popularity: a text mining approach. ProcediaComput. Sci. 92 (2016), 468–474.

[7]. T. K. Das, D. P. Acharjya, and M. R. Patra, “Opinion mining about a product by analyzing public tweets in Twitter,” 2014 Int. Conf. Comput. Commun. Informatics Ushering Technol. Tomorrow, Today, ICCCI 2014, pp. 3–6, 2014.

[8]. S. H. Doong, “Predicting twitter hashtags popularity level,” Proc. Annu. Hawaii Int. Conf. Syst. Sci., vol. 2016–March, pp. 1959– 1968, 2016.

[9]. Pang B, Lee L, Vaithyanathan S (2002) Thumbs up?: Sentiment classification using machine learning techniques. In: Proceedings

[10]. Wilson T, Wiebe J, Hoffmann P (2005) Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In: Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/ EMNLP 2005)

[11]. Johansson F, Brynielsson J, Quijano MN (2012)Estimating citizen alertness in crises using social media monitoring and analysis. In: Intelligence and Security Informatics Conference (EISIC). pp 189–196

[12]. Sun, S., Luo, C., Chen, J., 2017. A review of natural language processing techniques for opinion mining systems. Inf. Fusion 36 (2017), 10–25.

[13]. Thompson, J.J., Leung, B.H.M., Blair, M.R., Taboada, M., 2017. Sentiment analysis of player chat messaging in the video game StarCraft 2: Extending a lexicon-based model. Knowl.-Based Syst. 2017, 1–14.

[14]. Pak and P. Paroubek. „Twitter as a Corpus for Sentiment Analysis and Opinion Mining." In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320-1326

[15]. R. Parikh and M. Movassate, “Sentiment Analysis of User- GeneratedTwitter Updates using Various Classi_cation Techniques," CS224N Final Report, 2009

[16]. Bifet and E. Frank, "Sentiment Knowledge Discovery inTwitter Streaming Data," In Proceedings of the 13th InternationalConference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1-15.

[17]. AndranikTumasjan, Timm O. Sprenger, Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.

[18]. Yates, D., Paquette, S., 2011. Emergency knowledge management and social media technologies: a case study of the 2010 Haitian earthquake. Int. J. Inf. Manage. 31(1), 6–13.

[19]. Kim, J., Hastak, M., 2018. Social network analysis: characteristics of online social networks after a disaster. Int. J. Inf. Manage. 38 (1), 86–96.

[20]. Rayees Ahmad, Yasmin Shaikh, “Opinion Mining and Sentiment Analysis for Classification of Opinions on Social Networking Sites Using Machine Learning Algorithms: Systematic Literature Review” IJARCCE, Vol. 10, Issue 5, May 2021

[21]. Wang, Hao, et al., 2012. A system for real-time Twitter sentiment analysis of 2012 us presidential election cycle. In: Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics.

[22]. J. Roesslein. (May 2009). Tweepy Documentation. [Online]. Available: http://tweepy.readthedocs.io/en/v3

[23]. J. Akaichi, Z. Dhouioui, and M.J. Lopez-Huertas Perez, "Text mining facebook status updates for sentiment classification," in System Theory, Control and Computing (ICSTCC), 2013 17th International Conference, Sinaia, 2013, pp. 640- 645.

[24]. Danneman, N., Heimann, R., 2014. Social Media Mining with R: Deploy CuttingEdge Sentiment Analysis Techniques to Real-World Social Media Data Using R, www.it-ebooks.info.

[25]. F.Rustam,A.Mehmood,M.Ahmad,S.Ullah,D.M.Khan,andG. S. Choi, ‘‘Classification of shopify app user reviews using novel multi text features,’’ *IEEE Access*, vol. 8, pp. 30234–30244, 2020.

[26]. R. Khan, F. Rustam, K. Kanwal, A. Mehmood, and G. S. Choi, ‘‘U.S. based COVID-19 tweets sentiment analysis using TextBlob and supervised machine learning algorithms,’’ in *Proc. Int. Conf. Artif. Intell. (ICAI)*, A