LIVE TWITTER SENTIMENT ANALYSIS USING STREAMLIT FRAMEWORK

[1]Shilpa Patil, [2]V Lokesha

[1]shilpavskvsc@gmail.com,[2]v.lokesha@gmail.com

[1]Research Scholar, Department of Studies in Computer Science, Vijayanagara SriKrishnadevaraya University, Ballari, Karnataka –India.

[2] Department of Studies in Mathematics, Vijayanagara SriKrishnadevaraya University, Ballari, Karnataka –India.

Abstract

Sentiment analysis as per the textual description suggests that, it provides the sentiments or emotions on any form of given data in real life. As the social media information, now-a-days is flooded with various kinds of data from Facebook, Instagram, Twitter, and Whatsapp and so on. It is the need to handle data intelligently and classify as malicious and genuine data. The resultant growth in the area of social media entices tremendous challenges to the researchers in the field of social media analytics and deep learning. This paper contributes a user-friendly web application on 'sentiment analysis of live twitter data' using the keyword or handle, built on TextBlob library available in Python and Streamlit framework. The input data on the web application is firstly preprocessed with data cleaning, feature extraction and unstructured dataset view is filtered. Pre-processed data is further analyzed to collect the sentiments from a given twitter posts and predict into three classes viz positive, negative and neutral. The results presented in the work highlights the analysis of sentiments extracted from live tweets using keyword "Russian Ukraine World War" and classify the opinions as true or false tweets with positive, negative or neutral opinion by the use of Cat Boost Classifier.

Keywords: Sentiment Analysis, Streamlit, framework, Cat Boost Classifier, Textblob library.

I. <u>Introduction</u>

Sentiment Analysis is a technique used for emotional labels during specific configuration of components. It is one of trending observation of emotions that helps us to express the people's behavior either 'Positive', 'Negative' and 'Neutral'. For example, physiologically, an emotional label names such as 'Scary' and 'Happy' which are inclined with the increase or decrease of the blood flow of person. Sentiment analysis is defined as a task of Natural Language Processing (NLP) and Information Extraction. The Information extracted on Web during opinion collection involves data in the form of words, sentences and documents associated with various polarities. The primary task of sentiment analysis is to determine the polarity with classification for a given word, phrase, sentence and documents etc. There are varied numbers of challenge in the area of sentiment analysis. At one state/ locality is considered to be positive or in other state to be negative as first challenge. A second challenge is that the people never may or may not express same opinions. For example, however, “I Like Music” is more varying from “I am not Fond of music”. People are in contrast with their opinions by giving both positive and negative ratings.

A. Related works - Classification of Sentiment Analysis

As Bing Lui et al., stated that, Sentiment analysis has three main levels which are sub-categorized based on the techniques used, dataset structure and rating level etc. Again, each category is sub-categorized as shown in below fig1.

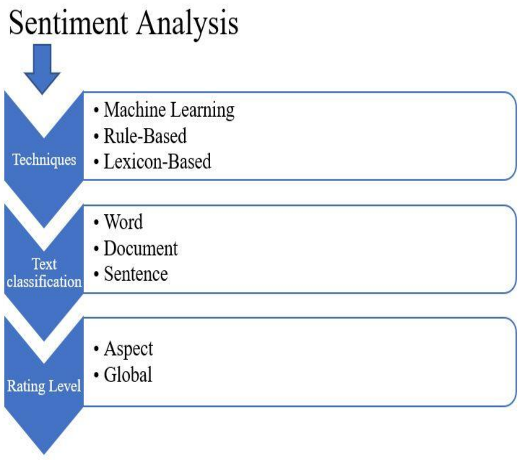


Fig 1: Categorization of Sentiment Analysis

Different Category of Sentiment Analysis

Machine learning approaches are used to analyze the data automatically. This means that a large amount of data collected in just few minutes, finds out the most positive and negative sentiments and then extracted the most important data for training on some specific criteria. ***Rule based*** systems is to establish a pattern for each tag, which rely on lexical analysis to interpret the input data with set of rules.

Lexicon-based method uses lexical analysis of sentiments which is to determine the polarity score for a given context. These represents a list of words with associated sentiment polarity score.

Text level analysis classifies into document and sentence-level sentiment analysis. At the document-level sentiment analysis classifies an opinion of a document with positive or negative emotions whereas the sentence-level tends to classify the opinions expressed in each sentence are pointed out either positive, negative or neutral. Hence, the basic difference of these two levels is not just provided the necessary opinion details on all perspectives.

Aspect and Global approach measure the strength of emotions for different aspects of a product and methods to review the rating on a global level that considers the polarity either as positive or negative.

Kaila et al., [2020] [8] proposed work on “Informational flow on Twitter-Corona virus outbreak - Topic Modeling Approach” related to flow of data on Twitter during the Pandemic of corona virus. The related tweets were analyzed using Sentiment analysis and using post preprocessing of Latent Dirichlet Allocation (LDA). Conclude that, when compared with earlier outbreaks of Ebola and Zika virus, there has been minimal misinformation with accurate and reliable tweets. Twitter API also made misinformation to be stopped and deleted immediately. Governments, Health and Organizations like WHO can be used on Twitter for spreading information and controlling panic among people outbreaks of virus. **Pillay et al., [2021] [14]** presented the paper on “Identifying emotions during Covid-19 using Topic Modelling Approach” identifies emotions in large population under unique situations like COVID-19. They concluded that “Topic Modelling Approach” or LDA methodology are used for twitter hashtags analysis with equal reliability and self-reporting studies. It also has a quicker process with a large number of respondents that gives better reliable data. Using Topic Model Approach, one aspect of measurement may serve as a proxy for others. Accessing emotions and using various measures to access each provided the most informative strategies. **Kaila et al., [2016] [2]** a paper on “An Empirical Text Mining Analysis of Fort McMurray Wildfire Disaster Twitter Communication using Topic Model”. Analyzed of wildfire disaster tweets on web-based communication i.e., on twitter, before, during and after disaster communication was done with most relevant, reliable and constantly updated information that helped stakeholders like residents, victims, governments, firefighters and other disaster management. Many tweets are downloaded and analyzed using frequency analysis, correlations analysis and Topic Model Latent Dirichlet Allocation. **Sanketh et al., [2020] [7]** paper on “Sentiment analysis of Live Twitter Data using Apache Spark” that handles different sources and different formats of structured and unstructured data, with the help of “Apache Spark framework”, which utilizes distributed memory abstraction. This Sentiment tweets are analyzed using our model and the results are displayed on visualization. There are multiple features like filters for selecting tweets that contains only hashtags of their interest. It also analyses positive and negative tweet score. Further, tweets are then stored and passed into the training data, to increase the accuracy of the model. **Alsaeedi et al., [2019] [6]** presented the paper “A Study on Sentiment analysis Techniques of Twitter data” explores various investigations of twitter data that have much attention over last decade with their outcomes. The paper presents various Machine Learning techniques, ensemble approaches and dictionary(lexicon)based approaches. Hybrid and ensemble Twitter Sentiment analysis techniques were also explored. Machine Learning algorithms such as Naïve Bayes, Maximum Entropy and Support Vector Machine, achieves an accuracy of approximately 80% using n-gram and bi-gram model. Ensemble and hybrid-based is better than Supervised Machine Learning Techniques, as they achieve accuracy of approximately 85%. **Jagdale et al., [2016] [3]** paper on “Sentiment analysis of events from Twitter using Open Source Tool” elaborates various methods of Sentiment analysis and focuses on the usage of datasets, to find the best approach respectively. Twitter API using R tool is used for collecting and preprocessing task. There are different approaches supervised and unsupervised, lexicon, dictionary and corpus-based methods that help us in sentiment analysis. Different dataset available for movie review, opinions dataset etc. In this methods, sentiment score (positive, negative and neutral) are calculated and counted number of tweets for given #Hashtags and can foresee opinions of public on a unique events. **Sultana et al., [2019] [1]** This paper represents on “Sentiment Analysis using Product review data” involves classification of text which analyses data and labels ‘better’ and ‘worse’ sentiment as positive and negative respectively. There are some challenges to find the accurate polarity of data using six well-known supervised classifiers. And noticed that the concatenations of adjectives, adverb and verb are the best combinations among various mixture of parts of speech. Visualizations are shown for outcomes of all combinations. **Elbagir et al., [2019] [4]** A study on “Twitter sentiment analysis using Natural language Toolkit and VADER sentiment”, “Valence Aware Dictionary for Sentiment Reasoner “(VADER) is to classify the text in Twitter data using multi-classification system. Tweet data are related to case study of US Presidential Election, 2016. The results obtained are of good accuracy in detecting ternary and multiple classes. Future scope presented in their work discusses on training large amount of data with lexicon and corpus for the better results. **Priyanka [2021] [9]** The proposed paper is on “Twitter Sentiment Analysis” addresses the problem of sentiment classification on twitter dataset and used number of machine learning and deep learning methods. For training and testing dataset used Anaconda distribution of Python for dataset with library requirements specific to some methods such as Keras with TensorFlow backend for Logistic Regression, MLP, RNN (LSTM) and CNN and XGBoost. For Preprocessing used, baseline, Naïve Bayes, Maximum Entropy, Decision trees, Random Forest, Multi-layer perception etc. are implemented. Neural networks is to classify the polarity of tweets with two types namely unigrams and bigrams, that increased accuracy of the feature vector with bigrams. This feature vector with bigrams. This feature may be extracted either Sparse Vector or Dense Vector. Sparse vector is recorded as better than frequency. In general, LSTM model achieves an accuracy of 83.0% on Kaggle whereas CNN model achieved 83.34%. Finally, taking the majority vote over prediction of 5 best models, with accuracy of 83.58%. **Rai et al., [2020] [10]** The paper is on “Sentiment Analysis of Twitter data” that gives detailed description of emotional analysis cycle that categorizes Twitters highly unstructured information into positive or negative. Several methods are used to destroy emotions including Twitter Knowledge-based strategies and Machine Learning Strategies. In addition, they have also presented Parametric correlation of strategies. As future opportunities in sentiment analysis includes improved techniques which can be applied to all data with less space. **Drus et al., [2019] [5]** This paper represents on “Sentiment Analysis in social media and it’s an application: Systematic Literature Review”, reports a systematic review of sentiment analysis in social media by exploring methods, social media platforms used and its application. The papers published between 2014 to 2019, have been reviewed using trustworthy and credible database that includes ACM, Emerald Insight, IEEE Xplore, Science direct and Scopus. The papers are reviewed based on the aim of the study. The result shows articles of sentiment in social media, extracted data on sites, which are mainly seen in world events, health Authorities, Entertainment, Politics and Business. **Ashique et al., [2021] [11]** The paper is on “Sentiment Analysis Using Machine Learning Approaches of Twitter Data and semantic analysis” is used to analyze data for a vast number of tweets of highly unstructured views either it is positive, negative or neutral. Firstly, pre-processed the dataset, then extracted and data with some context, that is termed as Feature Extraction. Machine Learning Classification Algorithms uses Naïve Bayes techniques is better than maximum entropy and the SVM subjects to Unigram model for betterment of results. Semantic Analysis on Wordnet, that increases the accuracy with these techniques. This shows the better visual representation for the users. **Golam et al., [2021] [12]** This paper proposed on “Investigation of different Machine Learning Algorithms to determine Human Sentiment using Twitter data”, using machine learning classifier algorithms which are applied on Crawled Twitter data with different types of preprocessors and encoding techniques, that end up with satisfactory accuracy. Later, achieved accuracies that has been showed. Evaluation of experiments shows that the Neural Network classifier algorithm provides accuracy of 81.33% with other classifiers.

Diyasa et al., [2021] [13] In this paper, tweets classifies with the keywords indihome, myindihome, useetv and wifi.di using textblob library and visualize the data using wordcloud. An accuracy value of sentiment analysis programs is calculated with best accuracy of 77.2%.

B. Problem in existing System

The Problem in existing sentiment analysis models is that they are all based and trained on structured datasets and open source, but are very outdated with time and most of the tools do not provide valuable information. Most of methods work on standard dataset available and are only limited to Open-Source datasets, that predict results only for that particular number of tweets. The analysis of the tweets is also limited in number. As per the literature reviewed, predicted result is not presented in case of live tweets and further only restricted number of tweets (20-55) were considered in the experimental study. Thus, resulting in the difficulty to analyze and predict the real time response of the users about any subject or topic.

II.Proposed work –Types of sentiment analysis

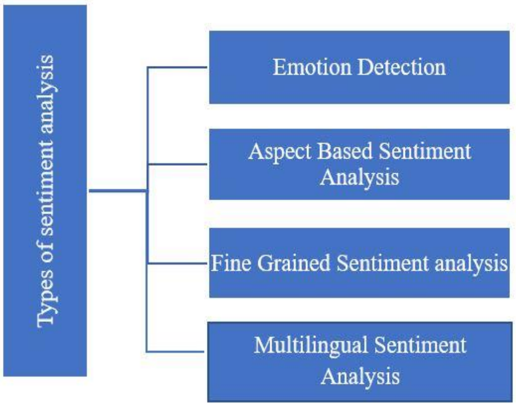

Types of sentiment analysis
Emotion Detection
Aspect Based Sentiment Analysis
Fine Grained Sentiment analysis
Multilingual Sentiment Analysis

Fig 2: Types of sentiment analysis

As mentioned in above fig2, there are different levels of analysis used data.

i. **Emotion detection:** It is the process of identifying human emotions that vary widely in their accuracy in multiple modalities such as automating the recognition of facial expressions from video, spoken expressions from audio, written expressions from text.

ii. **Aspect Based Sentiment Analysis:** As it breaks down the text into aspects (attributes or components), and the n allocates each one a sentiment level (positive, negative or neutral)

iii.**Fine Grained Sentiment Analysis:** Analyses the text and sentence level sentiments and classified the methods like stemming, bag of words, bi-gram, tri-gram and sentence level features are used to understand the sentiment polarity as very positive, positive, neutral, negative, very negative.

iv. **Multilingual Sentiment Analysis:** Sentiment analysis in multiple languages using complex neural network architectures with many pre trained models. Most popular are: Google’s BERT and XLNET, XLNET-2.

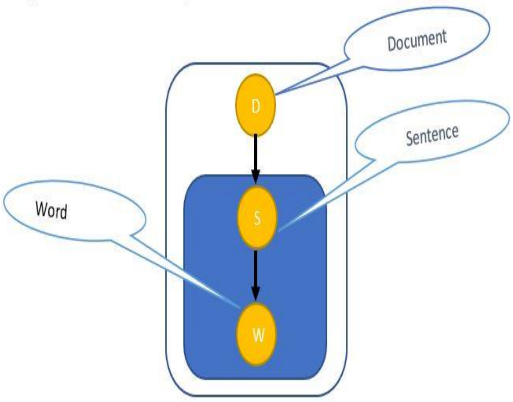


Fig 3: Levels of Sentiment analysis

Technically, social media was originated to communicate with friends and family-relations and later adopted by business intelligence as new communication method to reach out the users. The User activities include Photo Sharing, Blogging, social gaming, video conferencing, virtual world/gaming, reviews and much more. It is also used by governments and politicians for constituents and voters. Social media platforms are rapidly involved in huge amount of data which are being created among users. Many online networking companies like Twitter, Facebook, YouTube, Instagram and so on are exceptionally vary about notions and misinformation being spread on single entity. As data is growing with high extreme information in sentiment analysis to any kind of decision-making process either positive or negative, where millions of people have been communicated daily in social media platforms. Social media is also known as Internet-based where users can access quick electronic communication like personal issues /benefits, documents, videos and photos. Some social media applications are used to find network career opportunities, finding people across the world with like-minded interests and share their thoughts, feeling, insights and opinions.

III. Proposed method

The proposed architecture as shown in the figure 4 of live twitter sentiment analysis is based on the lexicon-based approach and involves training and testing phases for the classification problem in real time.

This model completely relies on querying the real time tweets from the Twitter API and pre-process them using Lexicon-based approach into a corpus of words. where the words are been predefined with a similar polarity score. This approach classifies the sentiment by counting number of positive and negative words after identifying the polarity score from the dictionary. Positive words depict the favorable conditions whereas negative words depict unfavorable conditions and last neutral depicts the neutral value ie zero. Lexicon-Based approach has 2 methods that is corpus-based methods and Dictionary based methods. Our project uses corpus based approach. As we have used Corpus based method to perform sentiment analysis in our project. Corpus method is complimented with Textblob for performing sentiment analysis for the model. Textblob is a library in python used for handling text data. It is a simple API used to perform basic NLP tasks such as WordtoVec model, POS tagging, stemming, Lemmatization, Named Entity Recognition, noun phrase extraction, translation, sentiment analysis using polarity score, classification, subjectivity score and so on. Textblob is based on NLTK library. From the fig5, it is understood that the

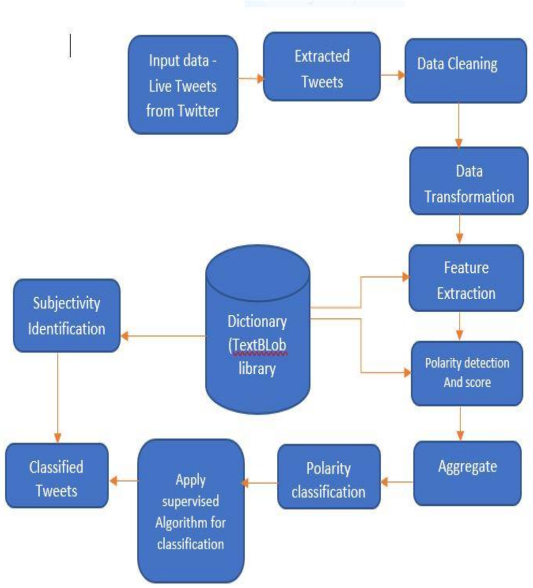

Input data - Live Tweets from Twitter
Extracted Tweets
Data Cleaning
Data Transformation
Feature Extraction
Dictionary (TextBLob library
Subjectivity Identification
Polarity detection And score
Aggregate
Polarity classification
Apply supervised Algorithm for classification
Classified Tweets

Fig 4: Detailed architecture of Live Twitter data

maximum used words in our tweets i.e. the Russian Ukraine World War dataset collected are survivor, killed shelling, Russia, BBC news and so on. When we explore more about the Russian Ukraine World War dataset, we can understand that many people are having neutral attitude and are optimistic about the situation, and extremely negative thoughts are more.

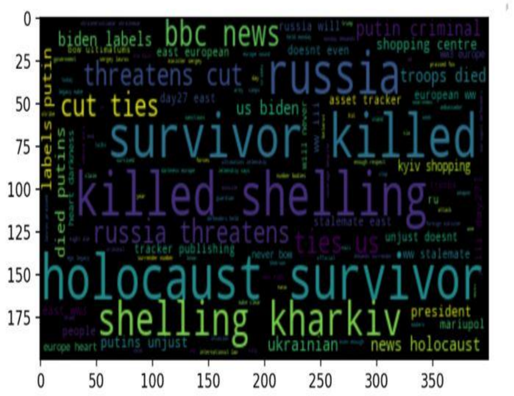
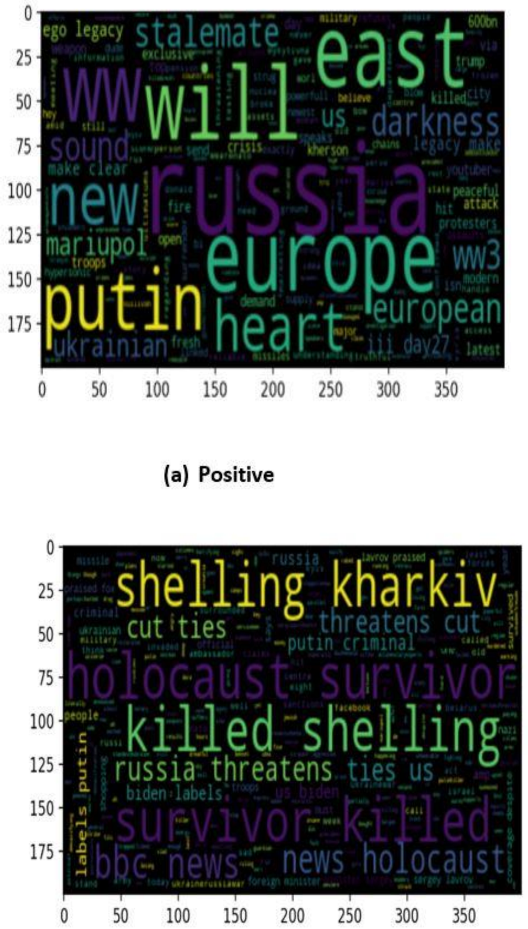


Fig 5: WordCloud depicts Frequent words



(a) Positive

(b) Negative

Fig 6: WordCloud depicts a) Positive b) Negative

IV. Data Classification – Cat Boost Classifier:

Machine learning has various applications in today’s world like transformation of handwritten digits to machine encoded traffic prediction, voice prediction, face prediction, biometrics, customer support and many more. Different machine learning techniques like Naïve Bayes, Logistic Regression, Random Forest, Support Vector Machine and Cat Boost and Gradient Descent classifiers are being significantly used by the researchers in the field of social media analytics. Few of the above machine learning classifiers are trained and transformed for the generation of word cloud on the topic “Russian Ukraine World War”.

Cat Boost Classifier It is open-source machine learning algorithm developed by Yandex researches. It is derived from 2 words Cat means category and Boost used for boosting machine learning algorithms. Typically used for gradient boosted decision trees. Cat Boost Classifier in Python inbuilt in ScikitLearn library that deals with categorical values automatically and can be easily integrated with deep learning algorithms as well. It does not require any data conversations like other data boosting algorithms.

It provide best accuracy without extensive data training used for both Python and R. After the dataset is being trained by different machine learning classifiers, we have summarized the accuracy of the classification report with respect to few algorithms in particular.

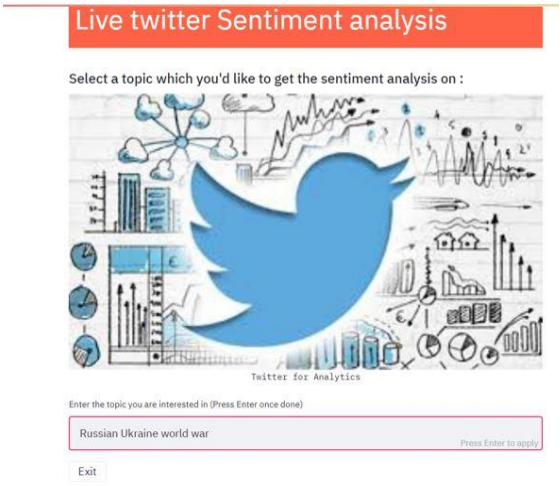
V. Data Visualization – Corpus-Based method

The data for the visualization process is collected from the real time tweets posted on the social media related to Russian Ukraine World War through Streamlit web API. Streamlit turns the data scripts into shareable web apps. Streamlit is one of the fastest way to build and share data apps. It's all about Python, open-source framework for building web apps for Machine Learning and Data Science, once we have created an instant web apps, we can use our cloud platform to deploy, manage and share the apps. Streamlit makes it incredibly easy to build interactive apps.

The sentiment analysis during the experimental study is carried for the first 201 Live Tweets on twitter as on dated 21/03/2022, and are categorized into three classes of sentiments such as Positive, Neutral and Negative.

VI. Results and Discussion

Snapshot1: At the beginning, enter the text/ topic of latest news.



Live twitter Sentiment analysis

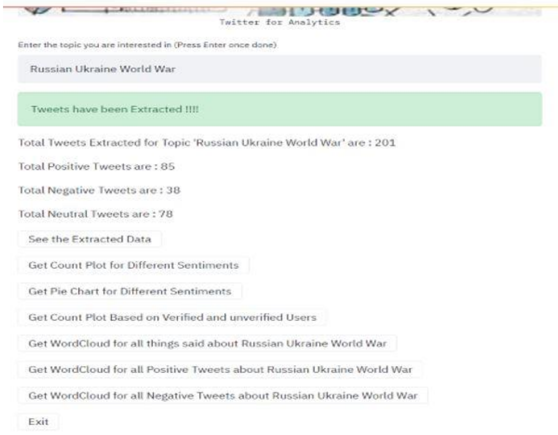
Select a topic which you'd like to get the sentiment analysis on :

Twitter for Analytics

Enter the topic you are interested in (Press Enter once done)

Russian Ukraine world war

Snapshot2: Click the options to generate extracted data, visualizing Pie chart and Bar chart, Word Cloud



Twitter for Analytics

Enter the topic you are interested in (Press Enter once done)

Russian Ukraine World War

Tweets have been Extracted !!!!

Total Tweets Extracted for Topic 'Russian Ukraine World War' are : 201

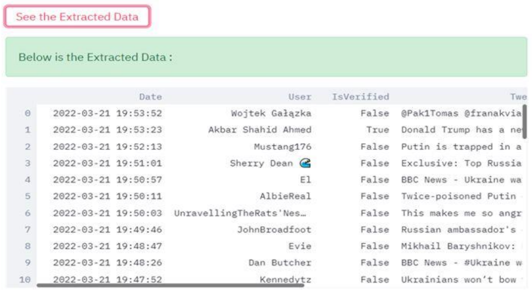
Total Positive Tweets are : 85

Total Negative Tweets are : 38

Total Neutral Tweets are : 78

See the Extracted Data
Get Count Plot for Different Sentiments
Get Pie Chart for Different Sentiments
Get Count Plot Based on Verified and unverified Users
Get WordCloud for all things said about Russian Ukraine World War
Get WordCloud for all Positive Tweets about Russian Ukraine World War
Get WordCloud for all Negative Tweets about Russian Ukraine World War
Exit

Snapshot 3: Extracted data with fields such as Date/ Time, User, Isverified, Tweet, Likes, RT, User location, clean tweet, sentiment.



Below is the Extracted Data :

	Date	User	IsVerified	Twe
0	2022-03-21 19:53:52	Wojtek Gałązka	False	@Pak1Tomas @franakvia
1	2022-03-21 19:53:23	Akbar Shahid Ahmed	True	Donald Trump has a ne
2	2022-03-21 19:52:13	Mustang176	False	Putin is trapped in a
3	2022-03-21 19:51:01	Sherry Dean 🌊	False	Exclusive: Top Russia
4	2022-03-21 19:50:57	El	False	BBC News - Ukraine wa
5	2022-03-21 19:50:11	AlbieReal	False	Twice-poisoned Putin
6	2022-03-21 19:50:03	UnravellingTheRats'Nes…	False	This makes me so angr
7	2022-03-21 19:49:46	JohnBroadfoot	False	Russian ambassador's
8	2022-03-21 19:48:47	Evie	False	Mikhail Baryshnikov:
9	2022-03-21 19:48:26	Dan Butcher	False	BBC News - #Ukraine w
10	2022-03-21 19:47:52	Kennedytz	False	Ukrainians won't bow

Snapshot 4: This is well depicted from the pie chart from fig7, for Russian Ukraine World war as it is observed that 16.92% of the tweets are positive,39.30% of the tweets are negative, 43.78% are neutral on 21-3-2022. These data shows that majority of public are having neutral opinion during this situation.

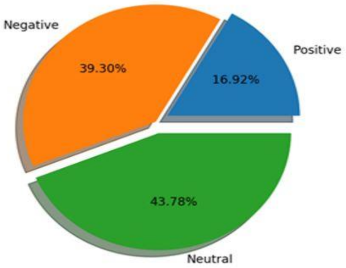
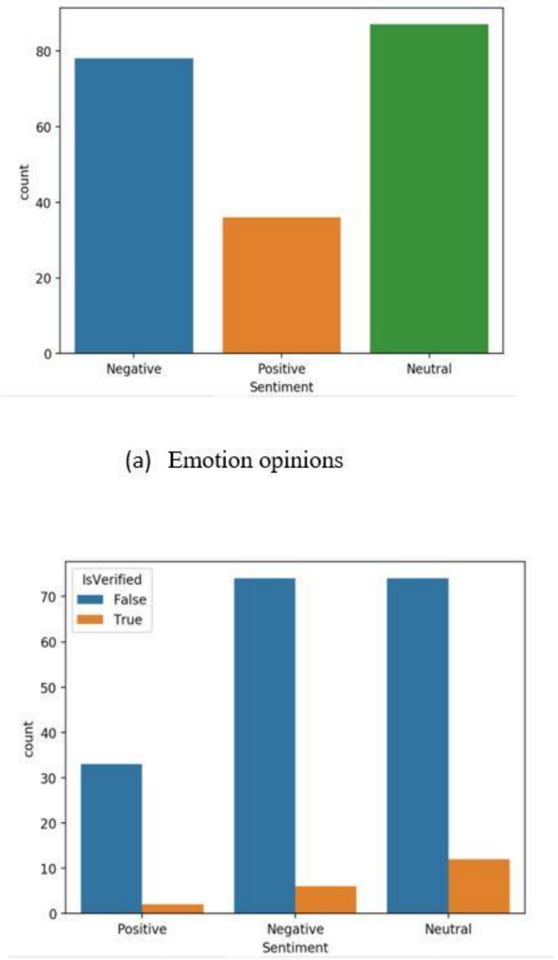


Fig 7: Pie chart depicting sentiments of live Russian Ukraine World War

Snapshot 5: Resultant analysis using Bar plots



(a) Emotion opinions

(b) True or false tweets

Fig 8: Results with Bar Plots

VII. Conclusion

The real time tweets considered for the experimental study were the mixture of hash-tags, symbols, URLs, words etc. The paper presents a supervised model for live tweets analysis and is subjected to series of preprocessing phases: Data Cleaning, Tokenization and Data Transformation before training phase. After testing multiple times, we observe that the system is able to detect sentiment polarity score of any given tweet correctly from the Twitter API for the first 201 extracted tweets. Results are visualized using BAR plot for better understanding on the topic "Russian Ukraine World War" and corpus of words generated are also disseminated with the frequency and occurrences of words. The experimental design is limited to English language and presently cannot identify sarcasm words. In future the model would be extended with robustness, to generate multi-lingual word cloud and support the sentiment analysis on broad social media platforms. The performance metrics: precision, recall, f1-Score, Cohen's Kappa for the real time sentiment analysis needs to be evaluated with various NLP specialized models.

References:

1. Fang, X., & Zhan, J. [2015]. Sentiment analysis using product review data. *Journal of Big Data, 2(1), 1-14.*
2. Prabhakar Kaila, D. [2016]. An Empirical Text mining analysis of Fort Mcmurray wildfire disaster twitter communication using topic model. *Disaster Advances, 9(7).*
3. Jagdale, R. S., Shirsat, V.S., & Deshmukh, S. N. [2016]. Sentiment analysis of events from Twitter using open-source tool. *IJCSMC, 5(4), 475-485*.
4. Elbagir, S., & Yang, J. [2019, March]. Twitter sentiment analysis using natural language toolkit and VADER sentiment. In *Proceedings of the international multiconference of engineers and computer scientists (Vol. 122, p. 16).*
5. Drus, Z., & Khalid, H. [2019]. Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science, 161, 707-714.*
6. Alsaeedi, A., & Khan, M. Z. [2019]. A study on sentiment analysis techniques of twitter data. *International Journal of Advanced Computer Science and Applications, 10(2), 361-374.*

7. Vemula sai Sanketh, Yashwanth Kumar Guntupalli, Devashish S Vaishnav [2020]. Sentiment analysis of Live Twitter Data using Apache Spark. *International Research Journal of Engineering and Technology (IRJET),7(8)*.

8. Prabhakar Kaila, D., & Prasad, D. A. [2020]. Informational flow on Twitter-Corona virus outbreak-topic modelling approach. *International Journal of Advances Research in Engineering and Technology (IJARET), 11(3)*.

9. Priyanka, V. [2021]. Twitter Sentiment Analysis (No.6065). Easychair.

10. Rai, S., S B, G., & Kumar, J. [2020]. Sentiment Analysis of Twitter Data. *International Research Journal on Advanced Science Hub, 2, 56-61*.

11. Ashique, M. [2021]. Sentiment Analysis Using Machine Learning Approaches of Twitter Data and Semantic analysis. *Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(6), 5181-5192*.

12. Golam Mostafa, I. A., & Junayed, M. S. [2021]. Investigation of Different Machine Learning Algorithms to Determine Human Sentiment Using Twitter Data. *International Journal of Information Technology and Computer Science (IJITCS), 13*(2), 38-48.

13. Diyasa, I. G. S. M., Mandenni, N. M. I. M., Fachrurrozi, M. I., Pradika, S. I., Manab, K. R. N., & Sasmita, N. R. [2021, May]. Twitter Sentiment Analysis as an Evaluation and Service Base On Python Textblob. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1125, No. 1, p. 012034). IOP Publishing.

14. Pillay, R. [2021]. Identifying Emotions During Covid-19 Using Topic Modelling Approach. *International Journal of Modern Agriculture, 10(2), 571-580*.