INFO 7390
ADVANCES IN DATA SCIENCE ARCHITECHTURES

UNDER GUIDANCE OF PROF RAM HARIHARAN

PROJECT TITLE:
SENTIMENT ANALYSIS ON TWITTER DATA

SUBMITTED BY

Nikita Gaurihar - 002980962
Ronak Patil – 001007035
Anshuman Karan - 001091491
Aishwarya Ghaytadak - 001568096

TABLE OF CONTENTS

Problem Statement

With the rapidly changing technologies, social media platforms like Facebook, Instagram, Twitter, etc. are becoming the data power sources for Data Scientists and Researchers. Nowadays, data has become an important aspect of running any business as it helps organizations to take data driven decisions through responsive insights, by performing trend analysis, predictive analytics, etc. To address and understand the real-world problems, businesses are looking for solutions with these disruptive technologies.

Like predictive analysis, we can perform trend analysis using the text data. Thus, in this project, we aim to predict the sentiments of the netizens who have tweeted on the Twitter platform in the recent period regarding the - "Ukraine-Russia war". Here, we are showcasing the polarity of the tweets and are inferring the overall sentiments about it.

Introduction of Sentiment Analysis

Sentiment analysis helps understand the performance of business through data visualization and understand several aspects of business operations such as customer satisfaction, analysis of demand and supply, understanding sales revenue, trends in media perceptions, Market Research, etc. This can be achieved by analyzing the comments, reviews and tone of the text that can be retrieved from different platforms such as LinkedIn, Amazon, Twitter, etc.

There are different types of sentiment analysis –

1. Standard Sentiment Analysis
2. Fine-Grained Sentiment Analysis
3. Emotion Detection
4. Aspect Based Sentiment Analysis
5. Intent Analysis

This project aims at analyzing the Twitter data about the most sensitive topic the world is hoping not to escalate further, i.e., Ukraine-Russia War. Here, in this project, we have followed the Fine-Grained Sentiment Analysis approach and predicted the sentiment polarity of the Tweets in the below mentioned Categories –

1. Highly Positive
2. Positive
3. Neutral
4. Negative
5. Highly Negative

Solution Approach

The detailed flow of this project is mentioned below -

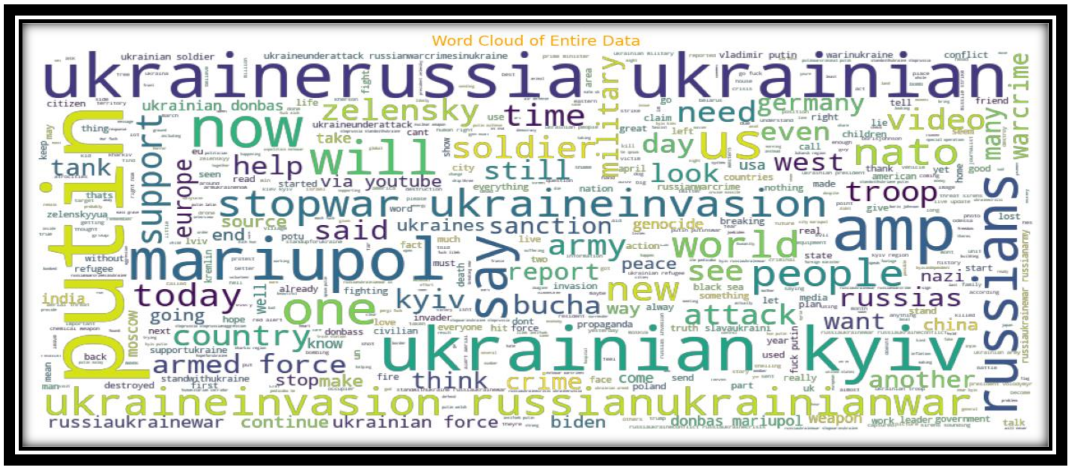
1. **Data Scrapping from Twitter:**
   - To extract data, we have used “snscrape”. Snscrape is a scraper for social networking services (SNS) that scrapes things like usernames, posts, hashtags, likes, etc. from a social media platform.
   - Snscrape helped us to scrape tweets from Twitter platform without the need of API.
   - We have scrapped around 164279 tweets with 7 attributes with following hashtags –
     1. #HopeForUkraine
     2. #savemariupol
     3. #StandWithUkraine
     4. #SupportUkraine
     5. #UkraineRussiaWar
     6. #zelensky
   - The list of attributes in our tweet extraction is mentioned below –
     1. Datetime
     2. Tweet Id
     3. Text
     4. Username
     5. Like Count
     6. Display Name
     7. Language

2. **Data Concatenation:**
   - We then collated data from 6 different extraction files in one dataframe using the python’s ‘concat’ function.
   - Considering this data as our final dataframe, we further performed Exploratory data analysis and data preprocessing using the Python’s Natural Language Tool kit – ‘nltk’.

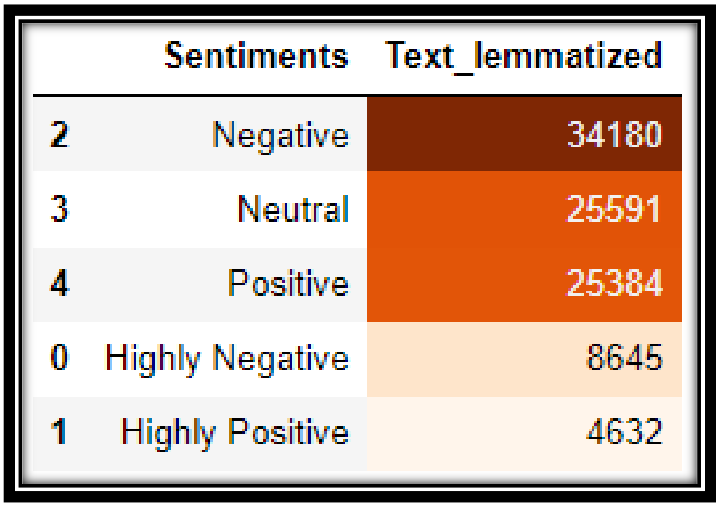
3. **Data Preprocessing:**
   - To perform Exploratory data analysis and data preprocessing using the Python’s Natural Language Tool kit – ‘nltk’.
   - Further, we cleaned the data, removed null values and performed the NLP preprocessing steps. The are mentioned as under –
     1. **Text_Clear** – Removed URL from data using Regex.
     2. **Text_Lower** – Converted our text data in lower case
     3. **Text_punctuation** – Removed all the symbols and punctuation marks
     4. **Text_Stop** – Removed stop words using ‘English’ library and customized the stop words list which was prevalent to our text data.
     5. **Text_Tokenized –** Split the sentences in list of comma-separated words
     6. **Text_Stemmetised –** Converted the words in its true form.

4. **Predicting Polarity of Tweets:**
   - ➢ Determined the polarity of tweets using the ‘Vader Sentiments’ python package.
   - ➢ Using the “SentimentIntensityAnalyser”, we segregated tweets with 5 polarities.

5. **Data Visualization:**
   - ➢ Following visualizations were made –

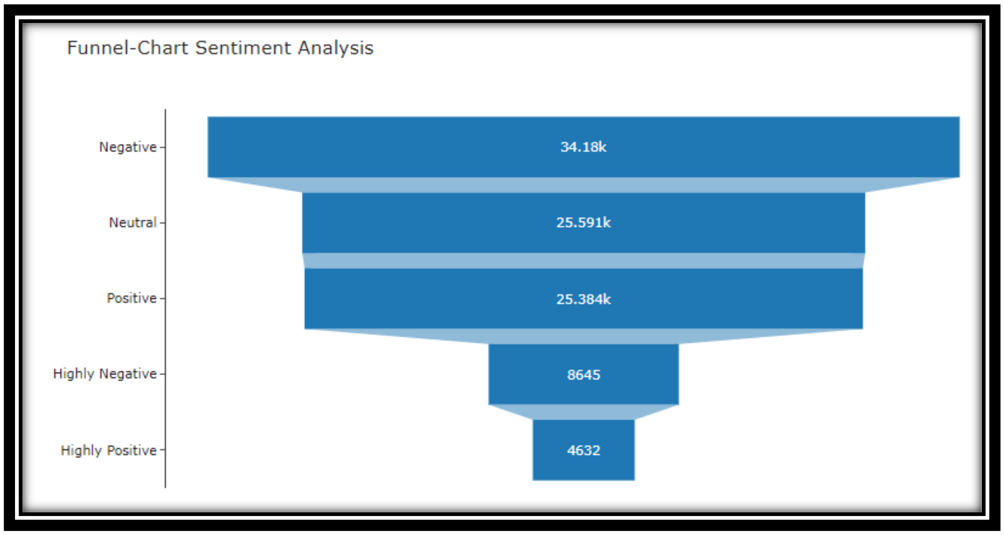


The following WordCloud displays the most frequently used words based on the total amount of tweets scraped from Twitter.

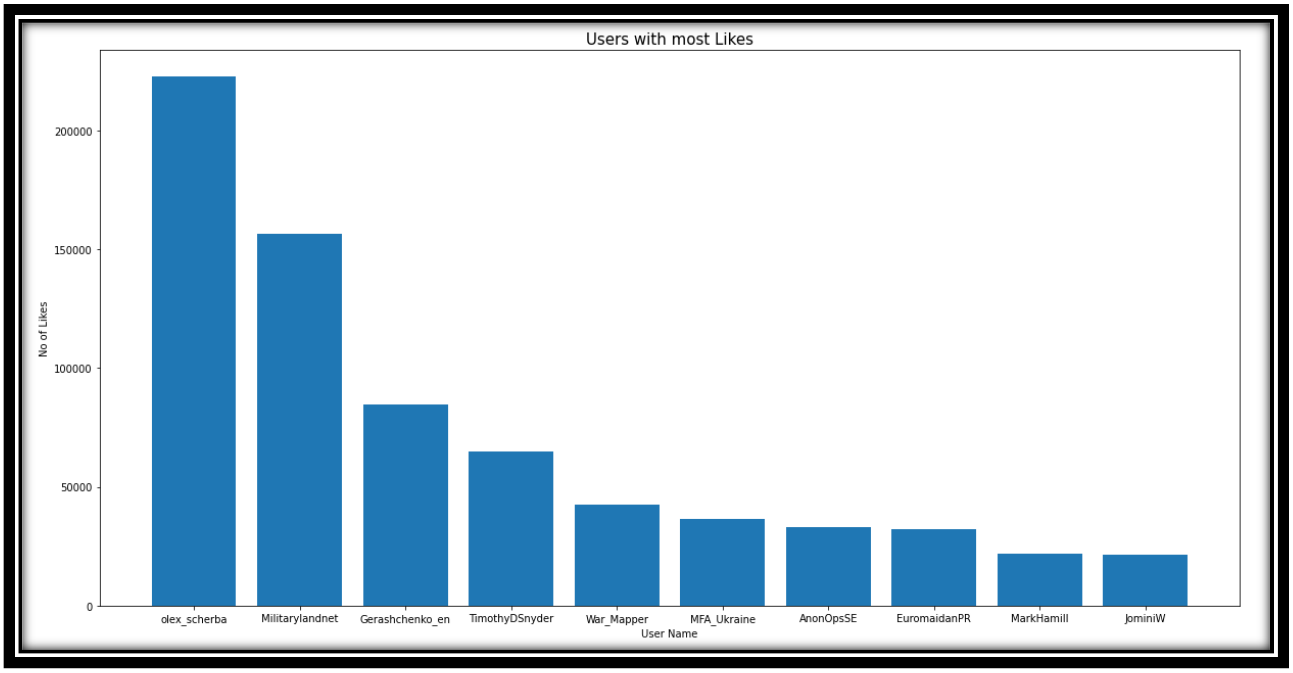


	Sentiments	Text_lemmatized
2	Negative	34180
3	Neutral	25591
4	Positive	25384
0	Highly Negative	8645
1	Highly Positive	4632

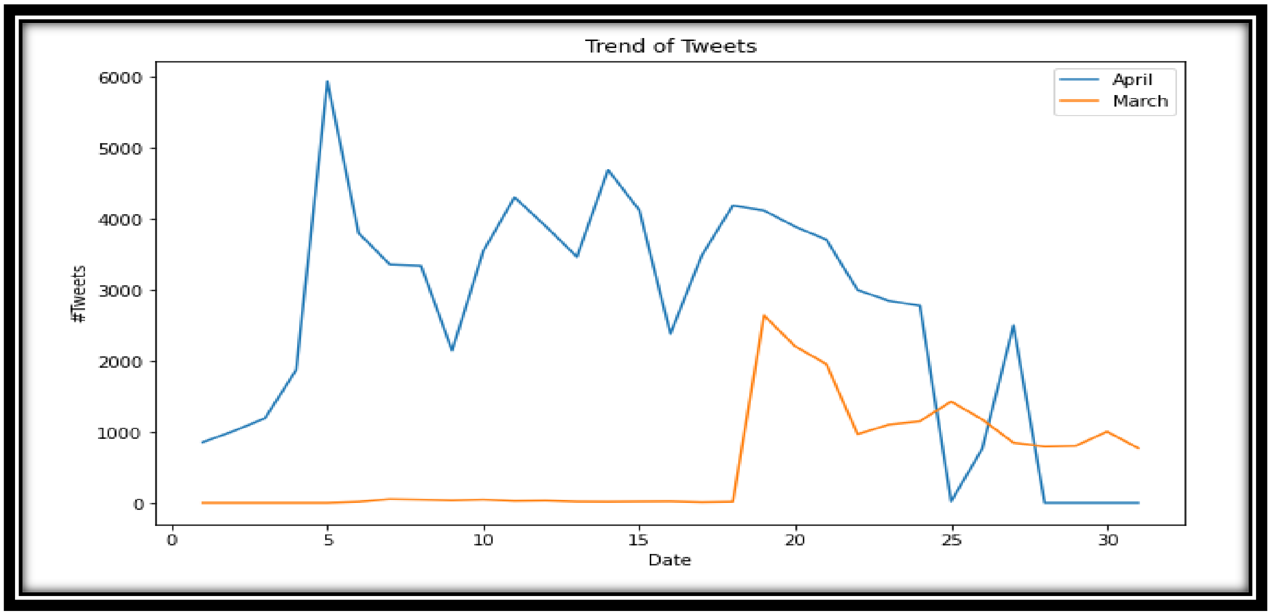
- ➢ The complete number of sentiment categories included in the dataset can be seen here.
- ➢ We can see that the majority of tweets have negative sentiments because the issue we chose had largely negative side effects.



- ➢ The funnel chart of all the sentiments in the database may be seen here.
- ➢ The counts of neutral and positive sentiments are nearly same, and we can see that there are some tweets with Extremely Negative and Extremely Positive sentiments.



We can see the most active Twitter users and their engagement in this particular topic using this visualization.



Here we can see the trend line of the tweets for the following months and the hype of the particular topic during that month.

6. Vectorization:

- Vectorization is a process used to convert the input data from its raw format in numeric form to support the ML Models.
- There are various vectorization techniques –
  1. Bag of Words
  2. Tf-Idf
  3. Word2Vec
- We have used 'TF-IDF' vectorization method to convert the text data in numeric form as it gives the words that occur more frequently in one document and less frequently in other documents and gives more importance to it since it is more useful for classification.

7. Prediction through Data Modeling

- Further, we split the data in train and test set with split ratio as 80:20.
- Performed 3 classification models namely – Logistic Regression, Random Forest Classifier and Decision Tree Classifier.
- Performed hyperparameter tuning to find the best parameter that gives better model accuracy.

8. Conclusion

- We have implemented 3 models such as Decision Tree Classifier, Random Forest Classifier and Logistic Regression Classifier and performed hyperparameter tuning to get best parameters

- Logistic Regression Classifier is the best model with 72% and accuracy of Grid Search CV model with 77%

References:

1. https://www.commsights.com/benefits-of-sentiment-analysis-for-businesses/
2. https://www.analyticsvidhya.com/blog/2017/01/ultimate-guide-to-understand-implement-natural-language-processing-codes-in-python/
3. https://machinelearningmastery.com/hyperparameter-optimization-with-random-search-and-grid-search/
4. https://github.com/rashidesai24/Analyzing-Twitter-Trends-On-COVID-19-Vaccinations