Twitter Analysis of the Russian Invasion of Ukraine

Social Media Mining : COMP-ENG-510

Devashri Naik, Rishita Jain

June 2022

1 Introduction

Russia Ukraine War began in February 2014 following the Ukrainian Revolution of Dignity, This war was primarily focused on the status of Crimea and the Donbas, internationally recognised as part of Ukraine. The conflict started increasing in late 2021 following a Russian military build-up on the Russia–Ukraine border. The intensity of the war increased exponentially when Russia invaded Ukraine on Feb 24, 2022. This war marked a strong change in views of people around the world. Suggesting a potential threat to Russia, they demanding Ukraine to not join NATO.

Ever since Russia launched a full-scale military invasion, the fighting has caused nearly three thousand civilian deaths and internally displaced more than seven million people, according to the United Nations. The conflict has forced another five million Ukrainians to flee to neighboring countries. Social media sites were filled with updates of the current situations typically the deaths, how the people were living and how the war was currently going on. People living in the neighbourhood shared photos videos and updates using different platforms and the people around the world shared their views regarding the war.

In light of the current scenarios of the Russian invasion, there are a lot of discussions that are observed on different social media platforms. Twitter is one of the platforms where different tweets, updates, and hashtags are shared that are trending in line with the war. Analyzing this data to get better insights into the current scenario has a lot of potential. Our problem statement is to analyze multiple tweets that are trending since the beginning of the war and incorporate different analyses on the extracted data.

2 Problem Statement

The Russian invasion on Ukraine caused a lot of discussion threads on Twitter with trending hashtags. People were expressing their opinions and situations. The information that these tweets contained could be used for multitude of things, hence we defined a few problem statements that we wanted to target in this project.

1. To perform text analysis and exploratory data analysis on the data. This will give a general trend in the data structure and how the distribution of data is with respect to different parameters.

2. To get insights from location of the tweets and try to understand the influence the war has on the region.

3. Finding the trending hashtags and analyzing the sentiments of the hottest hashtags.

4. To get an idea of the overall sentiment of the people's tweets, we performed sentiment analysis to understand the percentage of positive, negative and neutral tweets.

3 Data Set

In order to collect the data for the project, we crawled the web to get the Twitter data. After successful data extraction, we got few records but due to constraints on computational power and the number of tweets you can collect in a month for processing we considered using an open source dataset. The dataset that we considered is a Kaggle dataset. The same can be found at https://www.kaggle.com/datasets/bwandowando/ukraine-russian-crisis-twitter-dataset-1-2-m-rows.

This dataset was formed by a collection of tweets from the day the war started, i.e. Feb 24, 2022 and is updated everyday with the new file containing the daily tweets. For our simplicity we considered the tweets till May 24, 2022. Along the way, the author added a few columns but for data uniformity we kept the considered columns consistent. The dataset had the following columns: 'userid'-the id of the user, 'username' - name of the account, 'acctdesc' - account description, 'location' - the location of the tweet. The columns 'following', 'followers' which explains the number of people the user is following and the people who view the content shared by the user on daily basis. 'totaltweets' explains the number of tweets that the user created since it first started using the application, 'usercreatedts' explains the time when the user was first created. 'tweetid', 'tweetcreatedts', 'retweetcount' explains the information about the tweet like the unique tweet id, date and time when the tweet was created and the number of times the tweet was retweeted by others. 'text' explains the contents of the tweet which included the tweet in the origianal format along with the emojis. 'hashtags' were extracted separately in a column along with the language of the tweet in 'language' and 'coordinates'. There were two more columns explaining if the tweet was tagged as a favorite tweet in 'favorite count', and 'extractedts' explained when the tweets were extracted.

From all the columns above we decided to use 'userid' , 'username', 'location', 'tweetid', 'retweetcount', 'text', 'hashtags' for our analysis. We decided to discard the columns with date and time information as it was not relevant to our analysis. Also, the 'coordinates' column was majorly empty and 'favorite count' column was 0 always hence it was ideal to drop these columns. We will be explaining the analysis and usage of each column in the sections ahead.

3.1 Data Prepossessing

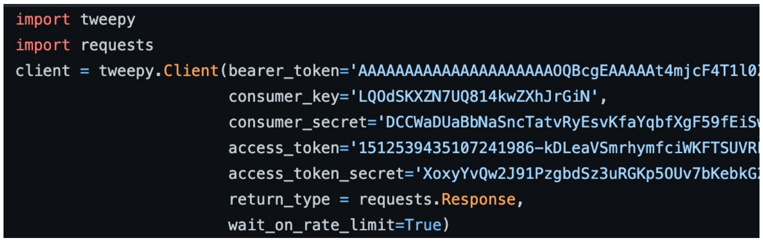
Twitter is a platform that not only contains huge amount of data but also makes it accessible. With the right analytical tools, this large amount of data can give us insights like user sentiments, spatial metrics etc. However, social media data is unstructured and needs to be cleaned before using. Most of the collected data contains miscellaneous information. The presence of URL's, emoticons and text case irregularities can hinder the process and give false insights. To avoid this, we pre-processed the twitter data used in the project. To increase the uniformity in the data, we performed the following process on the data:

1. Removal of duplicate tweets.
2. Converting all the letters to lowercase.
3. Analysis of English tweets for Sentiment analysis.
4. Removal of URLs and HTML artifacts (e.g., amp, \n), hashtags, mentions, digits, and emojis.
5. Removal of punctuation marks.

By pre-processing the data we could segment the data specific to the problem statements. This made analysis of data much simpler. Although only English tweets were used to analyse the sentiments, all language tweets were observed to get insights on the trends of the data and how different changes have affected the rise/fall of the tweets.

4 Technologies Implemented

Data cleaning and pre-processing helps streamline the data and makes the process of analytics easier. However, using the proper tools is as important. Below mentioned are the various technologies we have implemented in the project.



```
import tweepy
import requests
client = tweepy.Client(bearer_token='AAAAAAAAAAAAAAAAAAAAAOQBcgEAAAAAt4mjcF4T1l0
                       consumer_key='LQOdSKXZN7UQ814kwZXhJrGiN',
                       consumer_secret='DCCWaDUaBbNaSncTatvRyEsvKfaYqbfXgF59fEiS
                       access_token='1512539435107241986-kDLeaVSmrhymfciWKFTSUVR
                       access_token_secret='XoxyYvQw2J91PzgbdSz3uRGKp5OUv7bKebkG
                       return_type = requests.Response,
                       wait_on_rate_limit=True)
```

Figure 1: tweepy auth

4.1 Tweepy

Twitter is a useful source of text data; it has an API, credentials are easy to obtain, and there are a variety of Python packages available to assist with calls to Twitter's API. Tweepy is a simple Python package for interacting with this Twitter API. It makes accurate calls and is well-suited to dealing with tweets longer than 140 characters.

Tweepy supports both Basic Authentication and the newer OAuth technique for accessing Twitter. Because Twitter no longer accepts Basic Authentication, OAuth is the only way to access the Twitter API. Tweepy v2 makes the implementaion of OAuth easier. With tweepy, you can acquire any object and utilize every method that the official Twitter API provides. For example, a User object includes documentation, and tweepy may obtain the necessary information by following those recommendations.

One of the most common uses of tweepy is to observe for tweets and take action when an event occurs. The StreamListener object, which watches and captures tweets in real time, is a key component of this. StreamListener offers various methods, the most important of which are on data() and on status().

The new version of tweepy that is Tweepy v2 does not use the StreamListener object. With V2, the main difference is the authentication — instead of creating the API, you create the client with this code mentioned in figure 1.

4.2 Pandas

pandas is a Python library that provides quick, versatile, and expressive data structures that enable dealing with "relational" or "labeled" data simple and intuitive. It aspires to be the basic high-level building block for conducting realistic, real-world data analysis in Python. Furthermore, it aspires to be the most powerful and adaptable open source data analysis and manipulation tool accessible in any language. It is already well on its way to accomplishing this aim.

Here are a few examples of what Pandas excel at:

1. data (represented as NaN, NA, or NaT) in floating point and non-floating point data is easily handled.
2. Size mutability: columns in DataFrame and higher-dimensional objects can be inserted and removed.
3. Objects can be manually aligned to a set of labels, or the user can just disregard the labels and allow Series, DataFrame, and so on automatically align the data in calculations.
4. Powerful and adaptable group by capability for performing split-apply-combine operations on data sets for aggregation and transformation.
5. Allow for the simple conversion of ragged, differently-indexed data from various Python and NumPy data structures into DataFrame objects.

In the following project we implemented Pandas to create dataframes from the existing data. Overall four different dataframes were created, each consisting of tweets of one month. This made the data manageable. The process of exploratory data analysis became much easier.

4.3 BERTweet Model

BERT (Bidirectional Encoder Representations from Transformers) is a deep learning model in which every output element is linked to every input element and the weightings between them are dynamically determined depending on their relationship. BERTweet is the first public large-scale language model pre-trained for English Tweets. BERTweet is trained based on the RoBERTa pre-training procedure. The corpus used to pre-train BERTweet consists of 850M English Tweets (16B word tokens 80GB), containing 845M Tweets streamed from 01/2012 to 08/2019 and 5M Tweets related to the COVID-19 pandemic.

Sentiment was the biggest motivation in our project. Hence, we needed a model that could make the process manageable and easier for implementation. BERTweet made that plausible. For each tweet, the BERTweet model analysed if the tweet falls in the category of negative, neutral, or positive sentiments.

5 Analysis of Data

The structure of the data is on daily basis. Just observing a single day data will not help in analysis of changing people's sentiments. On contrary, although taking the entire available data can give a very high view generalized data, it is computationally heavy. As the data is divided on the basis of each day, to get the general trend analysis of data on monthly files made sense. For the analysis

Month	# Tweets per month	# days in war
February	735189	4
March	13618741	31
April	11107791	30
May	7861493	24

Table 1: Number of tweets corresponding to different months.

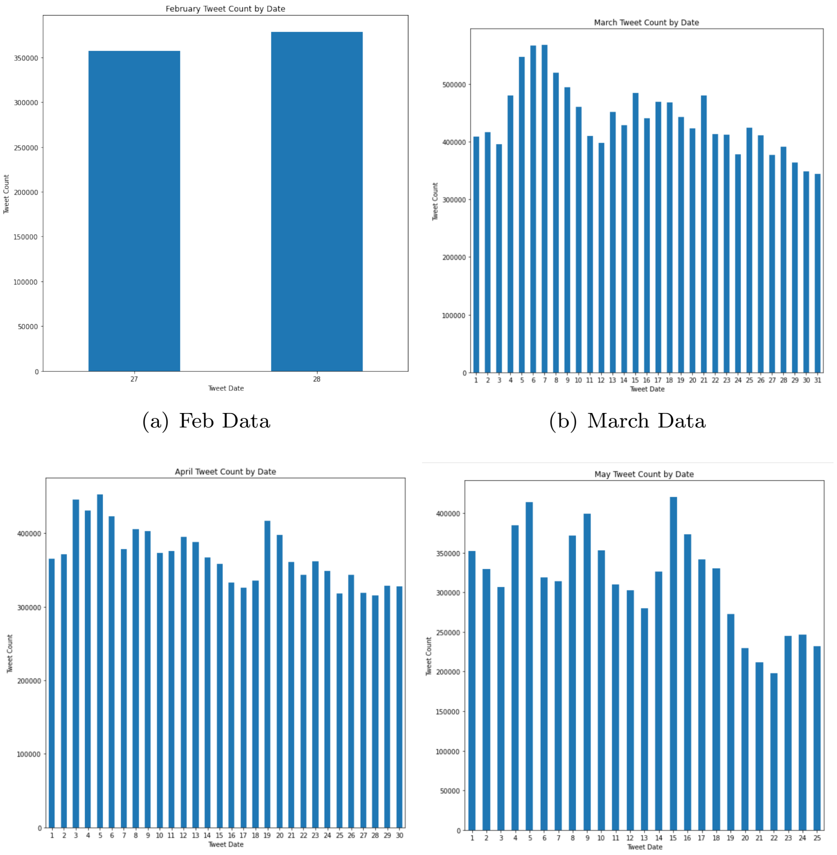
of data, we first started with performing basic exploratory data analysis. The data was divided into four data-frames, each denoting the tweets pertaining to a particular month. The EDA section shows results of both individual day tweet analysis and monthly tweet analysis. The Hashtag Analysis and Sentiment Analysis show result on monthly tweets.

Before diving deep into the data and understanding the nitty-gritty of the data, seeing the number of tweets on each day from the start of the war across different months was a good place to start. Table 1 gives a brief overview of the same.

5.1 EDA

As we observed above that the tweets for February were only observed for 4 days and there were about 350K tweets per day having an uniform distribution (a). This tells us that although the people were not as active in the later months where some days tweets even crossed 500K mark. The Fig 2. shows how the tweets per day for three months which approximately have 30 days each, i.e. March, April, May. As we can see from the figure (b) The number of tweets are highest from 4 March through 10 March. This can be because, people started getting more and more aware about the dire situation in Ukraine with the rise in the invasion. There were also videos surfacing how people there needed food and medicines to survive. Our hypothesis is that the surge in the tweets is because of these videos of invasion, and people asking for help. As seen from the April and May data in (c) and (d) respectively, we can conclude that most of the tweets per day in this topic ranged around 400K. Also, if we observe the late May days, the tweet count has significantly decreased, this could be because the topic became less trending with less people tweeting about it triggering more traction to this topic.

On observing the accounts with the highest tweet frequency we found that the top 20 users almost always remained the same with change in the sequence. Among them 'FuckPutinBot' always trends by a majority, which maybe because it is a bot account spamming the tweets. We wanted to analyze the frequency of English tweets with the tweets from other languages. As observed in the Fig 3. we can see that majority of tweets are in English Language. Very less tweets were done in Ukrainian language. We find out that on an average 66.80% of the



(a) Feb Data

(b) March Data

(c) April Data

(d) May Data

Figure 2: Tweets per day for all month

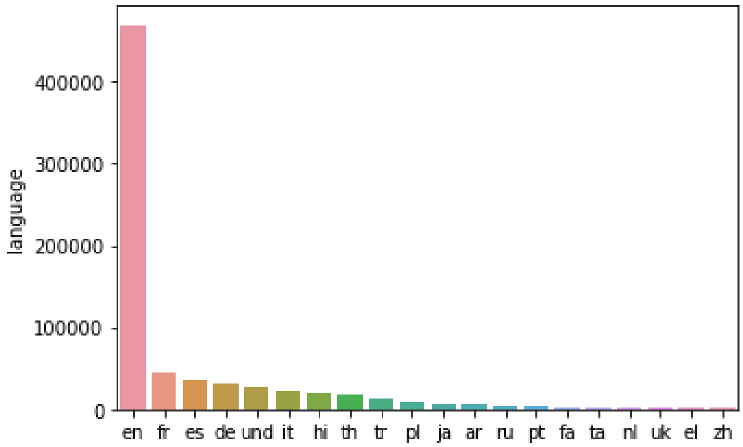


Figure 3: Language data distribution

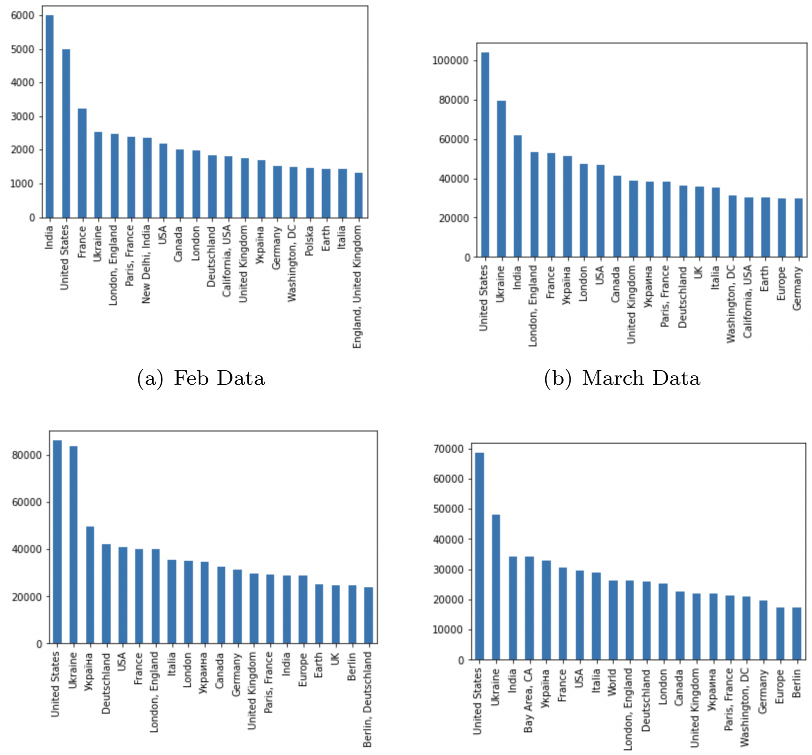
data is in English Language.

5.2 Location Specific Analysis

People on twitter do not always choose to put their location and getting this information specific to a tweet is difficult. This is why approximately 42.2% data is missing in our dataset on an average. This could cause a problem in the analysis. To get a location specific analysis, we considered only the tweets that had the 'location' data. As we can observe from subplots in Fig 4. (a) We can see that surprisingly majority of the tweets that had a geolocation were tagged in India. From all the other subplots, (b), (c) and (d) we can see that the highest geotagged tweets were shared by United States followed by Ukraine. The Figure 4 shows the top 20 countries over the four month span.

5.3 Hashtag Analysis

When one uses a hashtag in a Tweet, it becomes linked to all of the other Tweets that include it. Including a hashtag gives the Tweet context and allows people to easily follow topics that they're interested in. With tweets trending about the US, Hashtags in affiliation to 'ukraine','standwithukraine','russia','news' started trending. Tweets flooded the microblogging site with these hashtags making the people aware about the situation. These people started retweeting these tweets. Fig 6 shows the top 20 trending hashtags that are seen in 4 different months. Fig 5 shows the most retweeted tweet. As seen in that tweet the hashtags 'StandWithUkraine' is a thread that most people supported.



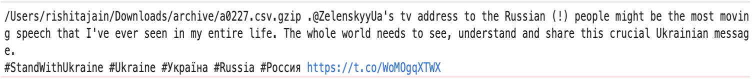
(a) Feb Data

(b) March Data

(c) April Data

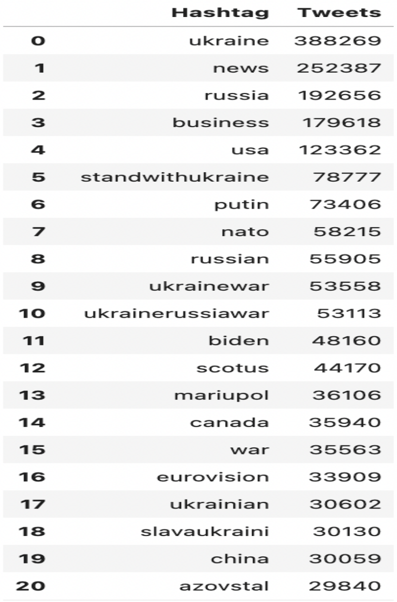
(d) May Data

Figure 4: Top locations for the tweets



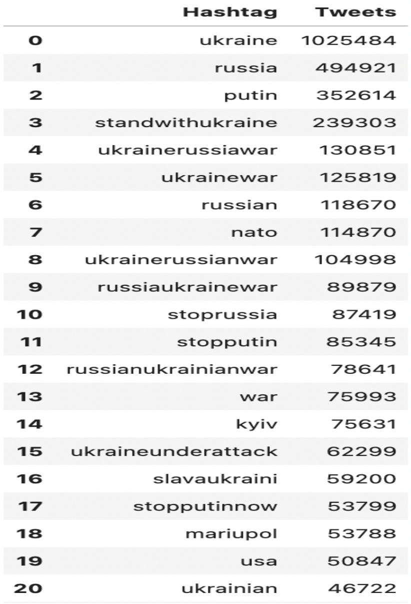
```
/Users/rishitajain/Downloads/archive/a0227.csv.gzip .@ZelenskyyUa's tv address to the Russian (!) people might be the most movin
g speech that I've ever seen in my entire life. The whole world needs to see, understand and share this crucial Ukrainian messag
e.
#StandWithUkraine #Ukraine #Україна #Russia #Россия https://t.co/WoMOgqXTWX
```

Figure 5: Most Re-Tweeted Tweet



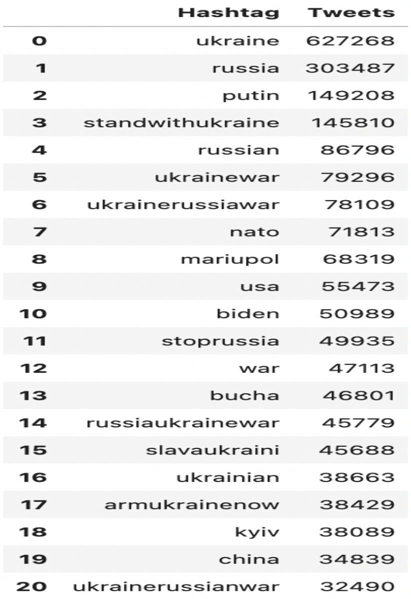
	Hashtag	Tweets
0	ukraine	388269
1	news	252387
2	russia	192656
3	business	179618
4	usa	123362
5	standwithukraine	78777
6	putin	73406
7	nato	58215
8	russian	55905
9	ukrainewar	53558
10	ukrainerussiawar	53113
11	biden	48160
12	scotus	44170
13	mariupol	36106
14	canada	35940
15	war	35563
16	eurovision	33909
17	ukrainian	30602
18	slavaukraini	30130
19	china	30059
20	azovstal	29840

(a) Feb Data



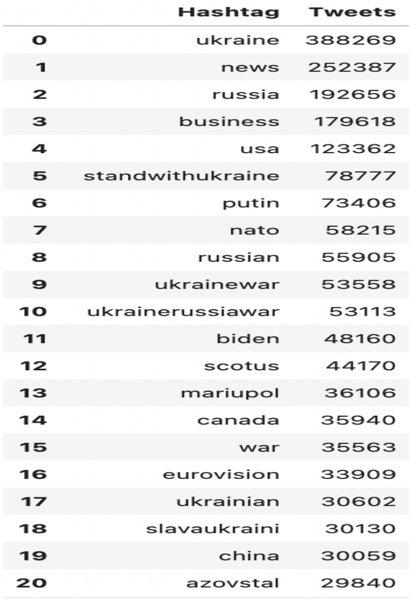
	Hashtag	Tweets
0	ukraine	1025484
1	russia	494921
2	putin	352614
3	standwithukraine	239303
4	ukrainerussiawar	130851
5	ukrainewar	125819
6	russian	118670
7	nato	114870
8	ukrainerussianwar	104998
9	russiaukrainewar	89879
10	stoprussia	87419
11	stopputin	85345
12	russianukrainianwar	78641
13	war	75993
14	kyiv	75631
15	ukraineunderattack	62299
16	slavaukraini	59200
17	stopputinnow	53799
18	mariupol	53788
19	usa	50847
20	ukrainian	46722

(b) March Data



	Hashtag	Tweets
0	ukraine	627268
1	russia	303487
2	putin	149208
3	standwithukraine	145810
4	russian	86796
5	ukrainewar	79296
6	ukrainerussiawar	78109
7	nato	71813
8	mariupol	68319
9	usa	55473
10	biden	50989
11	stoprussia	49935
12	war	47113
13	bucha	46801
14	russiaukrainewar	45779
15	slavaukraini	45688
16	ukrainian	38663
17	armukrainenow	38429
18	kyiv	38089
19	china	34839
20	ukrainerussianwar	32490

(c) April Data



	Hashtag	Tweets
0	ukraine	388269
1	news	252387
2	russia	192656
3	business	179618
4	usa	123362
5	standwithukraine	78777
6	putin	73406
7	nato	58215
8	russian	55905
9	ukrainewar	53558
10	ukrainerussiawar	53113
11	biden	48160
12	scotus	44170
13	mariupol	36106
14	canada	35940
15	war	35563
16	eurovision	33909
17	ukrainian	30602
18	slavaukraini	30130
19	china	30059
20	azovstal	29840

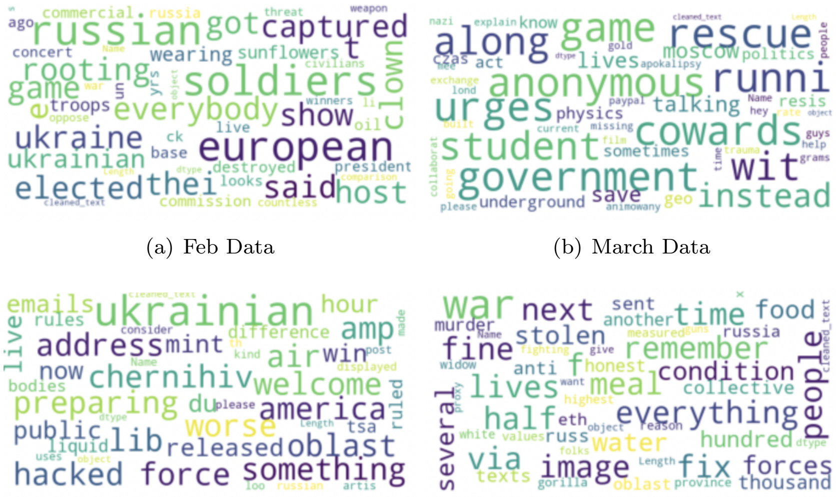
(d) May Data

Figure 6: Top 20 hashtags for each month

5.4 Sentiment Analysis

Before starting with the Sentiment Analysis, we can take a look at the WordCloud for different months. If we consider the data without removing the duplicate tweets, the word cloud is dominated by Ukraine, Russia, Putin, and StandWithUkraine. Hence, we decided to remove the duplicate tweets for the sake of creating WordCloud so that we can observe the different distribution of other more featured words as well by being less polarized. As we can see the word cloud is mixed and has no strong words that are prominent in all four clouds apart from ukraine. Although after taking a close look at the clouds, we can actually also observe the different sentiments of the people by the words used and the strength of the words shown by the times it is repeated. As we can see, cowards and government are dominant in March, Force, Hacked can be seen in April and steal, everything can be seen in May data. These clouds can be observed in Fig 7.

Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material. This helps understand social sentiment of a situation while monitoring online conversations. We divided the sentiments in Positive, Negative and Neutral Sentiments. As seen from the Fig 8, we can see that Negative emotions is far more noteworthy than Positive tweets. There are a lot of Neutral tweets as well. We suspect that this is because majority of the posts do not necessarily say bad things about the situation but expresses how the situation and how people can support Ukraine. Fig 9 is hashtags associated with texts segregated on the basis on sentiments from the texts.



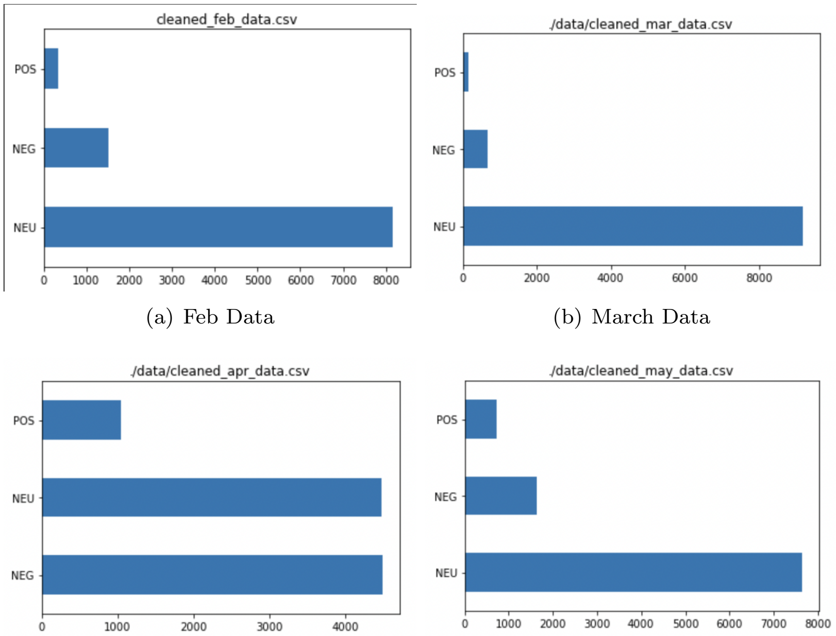
(a) Feb Data

(b) March Data

(c) April Data

(d) May Data

Figure 7: Word Clouds for every month



(a) Feb Data

(b) March Data

(c) April Data

(d) May Data

Figure 8: Sentiment Analysis for each month

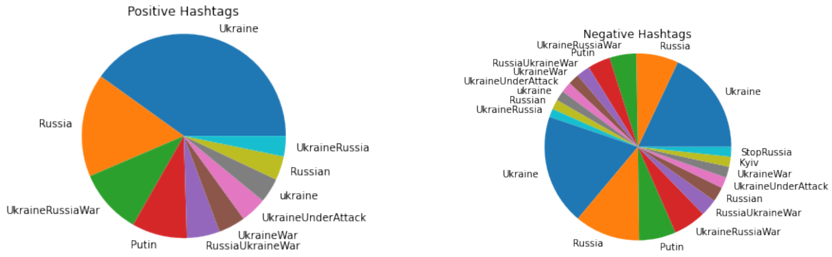
6 Discussion and Conclusion

The Russian invasion of Ukraine is atrocious act towards humanity. People have every right to express their displeasure and opinions on any platform they prefer. People need to stand with Ukraine. With this project we intended to analyse these emotions and get the general and the trend within tweets.

We performed text analysis and exploratory data analysis on the data. This gave us general trend in the data structure and how the distribution of data is with respect to different parameters. We see that the number of tweets were the most in the month of March and April and have been gradually decreasing. The most common tweet has been in english.

To understand the influence the war has on the region we analysed and plotted the location data to get insights from of the tweets. By performing Hashtag analysis, we were able to conclude that ukraine and standwithukraine is the most trending hashtag in the data.

With the results obtained from sentiment analysis, it was seen that most of the tweets skewed towards a neutral sentiment rather that a negative one.



(a) 20 Positive Hashtags

(b) 20 Negative Hashtags

Figure 9: Trending haghtags divided based on tentual sentiments

The only exception was the month of april, where the tweets skewed towards negative sentiment. Wr believe that it was due to the fact that in the month of april, A Russian rocket destroyed an airport runway in Odesa.

7 Challenges and Future Work

One of the main challenges was the sheer size of the dataset. With 35.95 million tweets within 85 csv files, one of the biggest issue was parsing the data into dataframes. Finding the right method for data segmentation and finally deciding on monthly division was arduous and time consuming.

One other challenge was the implementation of the tweepy v2 auth. Twitter recently changed the configuration of its API, making the use of Tweepy v2 authorization for tweet collection compulsory. Though the process in itself was easy, there weren't many resources available ultimately delaying the data collection process.

Finding the right method suitable for the sentiment analysis of our dataset was important. We first started the process with the RoBERTa Model, hoping it would be able to give better sentiment scores. However, the size of the dataset was an hindrance to the whole process. We ultimately decided on implementing the BERTweet Model. Even though we weren't able to get seniment scores, we believe that the BERTweet Model was able to perform better Sentiment Analysis.

There's a huge potential for future improvement in our project. We have performed sentiment analysis with our data set to garner insights on the general sentiment shared among the people in regards with the war. We believe that the implementation of Emotion Analysis would help us better perceive the thought process of the people with respect to such atrocity. We have implemented BERTweet Model for sentiment analysis, but we could incorporate other models like RoBERTa Model and XLM-R Model. These models could be compared and analysed for better sentiment scores.

8 References

1. https://www.kaggle.com/datasets/bwandowando/ukraine-russian-crisis-twitter-dataset-1-2-m-rows

2. https://www.kaggle.com/code/samupark/war-in-ukraine-feb-eda-and-sentiment-analysis

3. https://www.kaggle.com/code/ssaisuryateja/eda-and-sentiment-analysis

4. https://docs.tweepy.org/en/stable/client.html

5. https://dev.to/twitterdev/a-comprehensive-guide-for-using-the-twitter-api-v2-using-tweepy-in-python-15d9

6. https://www.kirenz.com/post/2021-12-10-twitter-api-v2-tweepy-and-pandas-in-python/twitter-api-v2-tweepy-and-pandas-in-python/

7. https://github.com/VinAIResearch/BERTweet

8. https://huggingface.co/docs/transformers/model_doc/bertweet