MadihaTanvir
Socialmedia&Webtechnology
Linnaeus University, Sweden
mt223rg@student.lnu.se

Introduction

Big data refers to huge, diverse, and complicated data collections that are challenging to store, analyse, and visualize for use in subsequent operations or outcomes. Big data analytics is the practice of analysing enormous amounts of data to uncover obscure patterns and hidden correlations. The effectiveness of social media has generated tremendous interest among Internet users nowadays.

These social networking sites data can be used for a variety of things, including sentiment analysis, marketing, and prediction. Twitter is widely used for posting comments through short status. Twitter has ended up an imperative stage which individuals are taking up to precise their views and conclusions approximately any topic. The millions of tweets received every year could be subjected to sentiment analysis.[1] Russia borders five EU member states and the EU surrounds the Russian exclave of Kaliningrad. Russia's biggest economic partner is the EU, and it plays a sizable share in the European energy industry. In order to cripple Russia's capacity to fund the conflict with Ukraine, put pressure on Russia's elites, and reduce Russia's economic foundation, the EU has implemented a number of rounds of increasingly harsh sanctions. Sanctions cannot be imposed unless all EU members agree to them.[2]

Hence this being a common topic in most of the social media platforms and on, when a user wants to share his opinion on a hot topic like this, the user tweets utilizing Hashtags, emojis, acronyms, and puns make it difficult to examine the data.

Approach to solving the given problem and tasks

Selection of Topic

Initially we choose Fitness as our topic for research and we sub categorized on yoga. After collecting data on yoga, we realized that it was itself a very vast topic for us as beginner in the field of big data. We changed the plan to work on Data Science we collected tweets in Data Science and after analysing the collected tweets we were again confused what to report on that data.

Finally, we choose the topic of Russia and Europe relationship and as Ukraine was already a topic of discussion so we exclude the Ukraine from the research so we can have idea without the specific impact of Ukraine we just collected the 1000 tweets on the search of Russia and Europe without giving any time as a parameter.We are opted to work on sentiment analysis on social media data on Twitter, to discover user attitudes or sentiment of the users.

Retrieving Tweets:

We create a tweeter developer account to get the consumer key and consumer secrete key as well as access key and access secrete key to fetch the data from the API. Following code is used to access the Data from API

```
# Twitter authentication
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_key, access_secret)

#  Creating an API object
api = tweepy.API(auth)
```

Collect tweets:

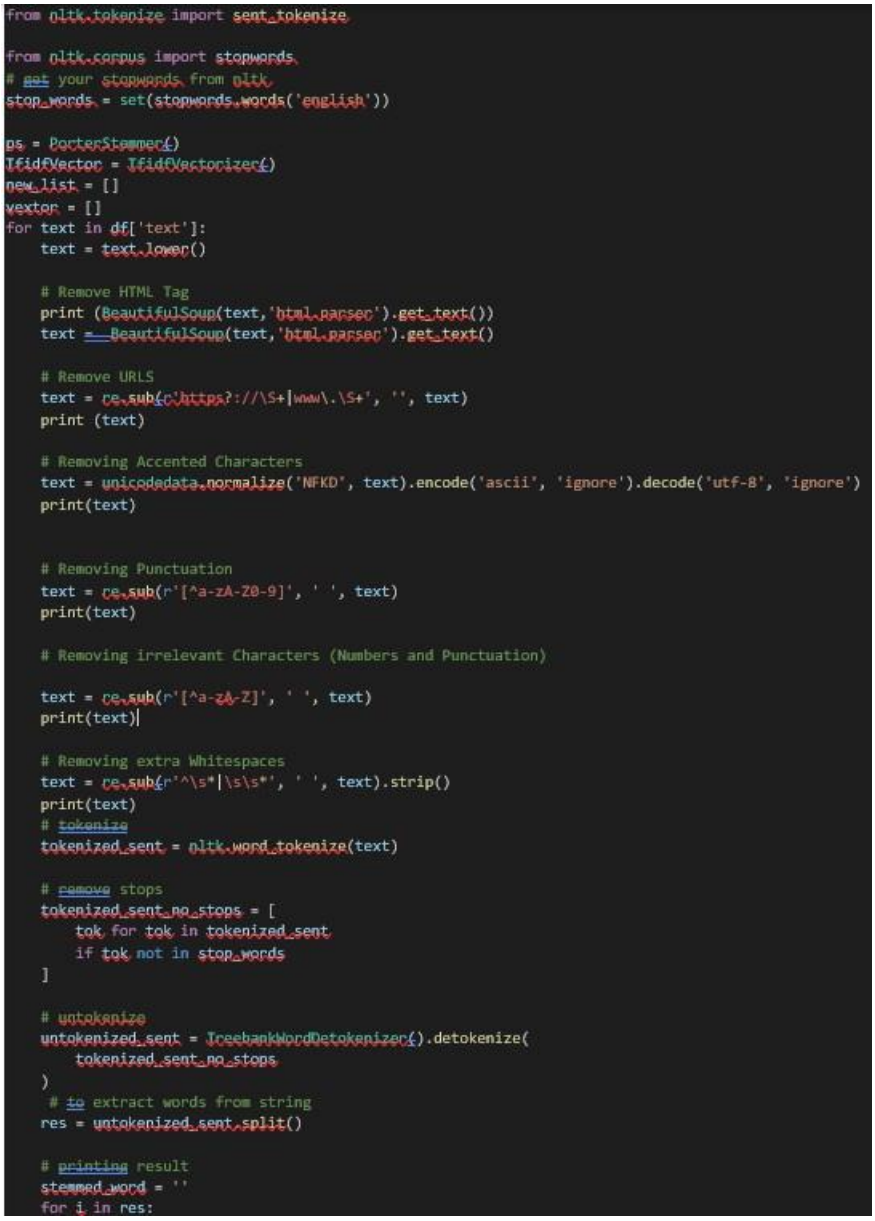
The Tweets about Russia and Europe was collected by using the following code. The tweet limit was set to 1000 and the language was selected as English only. Then the file was saved as CSV file. For the purpose panda library was used. File is attached in the Appendix

```
list = []
date_tweets = tweepy.Cursor(api.search_tweets, q='"Russia" "Europe"' , lang = 'en',
tweet_mode='extended').items(1000)

for tweet in date_tweets:
    text = tweet._json["full_text"]
    print(text)
    # preprocessing
    text = text.lower()
    text = re.sub(r'https?://\S+|www\.\S+', '', text)

    refined_tweet = {'text': text,
                     'favorite_count': tweet.favorite_count
                     ,
                     'retweet_count': tweet.retweet_count,
                     'created_at': tweet.created_at
                     }

    list.append(refined_tweet)




import pandas as pd
df = pd.DataFrame(list)
df.to_csv('Russia_Europe.csv')
```

Cleaning of Tweets:

The collected tweets were cleaned while using the NLTK libraries with these various functions including stop word removal, word tokenization and word stemming

Following code was used to clean the tweets



```
from nltk.tokenize import sent_tokenize

from nltk.corpus import stopwords
# set your stopwords from nltk
stop_words = set(stopwords.words('english'))

ps = PorterStemmer()
IfidfVector = IfidfVectorizer()
new_list = []
vextor = []
for text in df['text']:
    text = text.lower()

    # Remove HTML Tag
    print (BeautifulSoup(text,'html.parser').get_text())
    text = BeautifulSoup(text,'html.parser').get_text()

    # Remove URLS
    text = re.sub(r'https?://\S+|www\.\S+', '', text)
    print (text)

    # Removing Accented Characters
    text = unicodedata.normalize('NFKD', text).encode('ascii', 'ignore').decode('utf-8', 'ignore')
    print(text)


    # Removing Punctuation
    text = re.sub(r'[^a-zA-Z0-9]', ' ', text)
    print(text)

    # Removing irrelevant Characters (Numbers and Punctuation)

    text = re.sub(r'[^a-zA-Z]', ' ', text)
    print(text)

    # Removing extra Whitespaces
    text = re.sub(r'^\s*|\s\s*', ' ', text).strip()
    print(text)
    # tokenize
    tokenized_sent = nltk.word_tokenize(text)

    # remove stops
    tokenized_sent_no_stops = [
        tok for tok in tokenized_sent
        if tok not in stop_words
    ]

    # untokenize
    untokenized_sent = TreebankWordDetokenizer().detokenize(
        tokenized_sent_no_stops
    )
     # to extract words from string
    res = untokenized_sent.split()

    # printing result
    stemmed_word = ''
    for i in res:
```

Text feature extraction:

Text feature extraction is applied by using the Term Frequency Inverse document frequency library (TFidfVectorizer) using the sklearn module. Following code is used

Clustering while using K-Means algorithm:

Clustering is done by using K-Means Algorithm Following code was used to separately save all the 10 clusters into 10 different csv files.

```
from sklearn.cluster import KMeans
Vectorizer = TfidfVectorizer(stop_words = "english")
vector = []
all_keyword = []
new_text = ''
for text in df['0']:
    new_text +=text
new_text = [new_text]
print(new_text)
vector = Vectorizer.fit_transform(new_text)
features_name = Vectorizer.get_feature_names_out()
dense = vector.todense()
dense_list = dense.tolist()
```

Methodology approach:

A program is designed in Python and implemented to catch all the possible Tweets that match our requirement the topic Russia Europe. We used the Repl.it as our IDE. Library tweepy used for the import the data and Python code is used to integrate with the Twitter API for tweets extraction. We need to have consumer keys and secrets as well as access token key and secrets generated to use the API.[2] Then we access these tokens via our code. We extracted a list of tweets in a tab separated format as csv. The first number is tweet id followed by the tweet message.

Data Collection:

Multiple times each day, all tweets with the keyword phrase "Russia Europe" were methodically checked, and any repeated tweets were deleted. Only English-language tweets were included in the extraction. The retweets and replies were filtered out while collecting the tweets to avoid duplication of the tweets. As the complete database was obtained, the data cleaning process has been performed, where the asci code, URLs, html tags spaces, punctuation, stop words were removed. Then we extracted the words from the string and then we stemmed the words and created a new csv file as russiarefined tweets.

Sentimental analysis:

This stage involved identifying the feelings and sentiments expressed throughout the research social listening. A sentiment analysis system for text analysis incorporates natural language processing. For the sentimental analysis we used textblob and we created a function to assign the value of sentiments based on the polarity[6].

Code for the barchart

```
# Plot Bar Char
import matplotlib.pyplot as plt
plot = df['Sentiments'].value_counts().plot(kind= 'bar', title = 'Sentiment Analysis on Russia
Europe Relation')
```

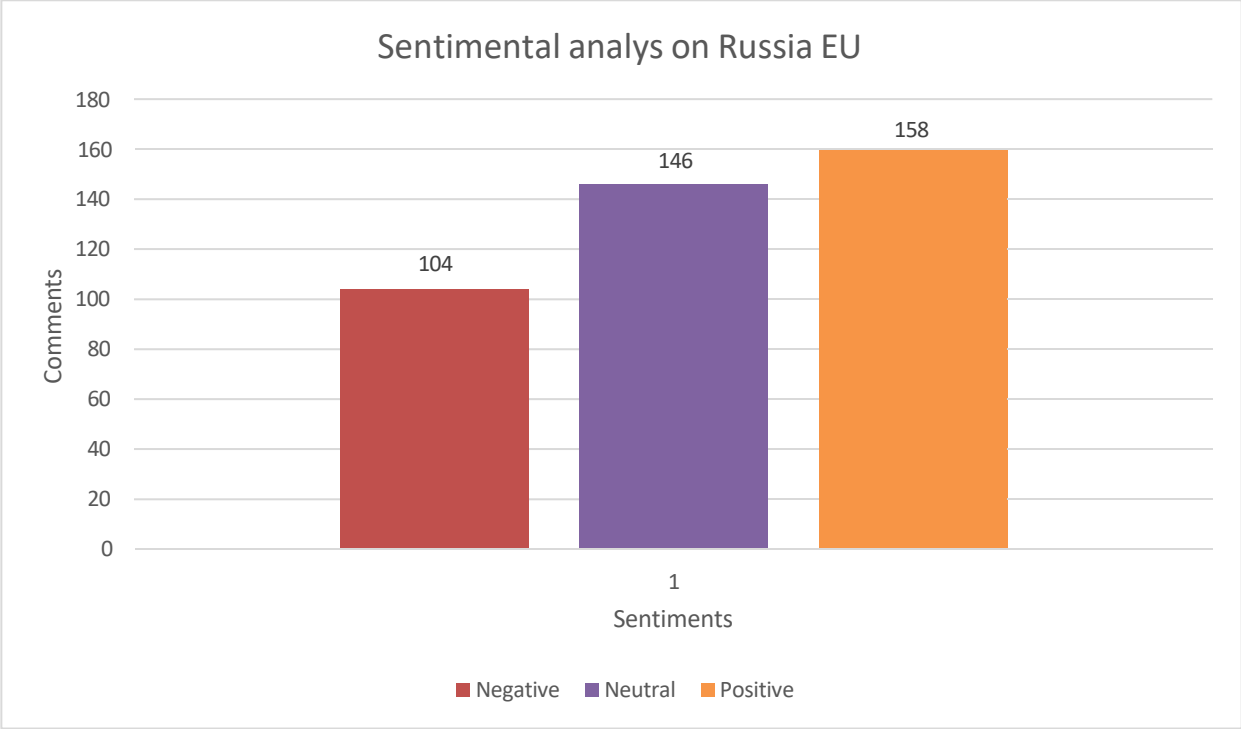


Fig:1 Bar chart for sentimental analysis

As everyone was talking on the issue of Ukraine with Russia the Topic on the Russia and Europe if taken from this year could have given us different results from our current results. We decided to exclude Ukraine from the research and we also exclude the timeline from the tweet fetching so we can get first 1000 tweets about the Russia and Europe and from the world cloud we can see that the Ukraine was a prominent topic of discussion (Fig.2) but still we get the positive sentiment analysis more than the negative.

The results could be different if we have collected more tweets and if we have excluded this year from the research.More work can be done to get a better understanding of the impact of Ukraine war on the sentiments of people about Russia and Europe.

Outcomes/Analysis of results

With the word cloud towards the most notable topics and concepts related to the keyword "Russia Europe" used in the downloaded tweets were determined by looking at the top 1000tweets, users talked about Russia, Europe, Ukraine, War, Putin, America, Attack, Zelensky and Weapon. Russia Ukraine conflict was reflecting in these tweets[4].

Fig 2: Word cloud of Russia Europe

Data as valuable resource:

Data is regarded as one of the most precious resources in the contemporary internet-centric environment we live in because of the potential money and economic value it may generate. Although Web 1.0 provided us with a fantastic means of connecting and communicating, we weren't really able to properly navigate through all the information at the time. Platform-based environments like Google, Facebook, and YouTube first appeared with the advent of Web 2.0. This facilitated data flow and access. From the data we gathered for this study, we can see how users are reacting to the term RussiaEurope . Since Twitter data has a structure in place that enables any user to follow another user and provides nearly all of its data through its APIs, Twitter is exceptional in this regard. And we could find a vast data on the topic we selected for this paper.

Consideration of data capturing:

While working with the data we realized that data should be captured from multiple social media tools instead of only one, in our report we get the data from Twitter only but by adding data from Facebook, YouTube and Instagram you can have a better result about the sentiments of people on our selected topic. Since the capture stage gathers data from many users and sources, a sizeable portion may be noisy and thus have to be removed prior to meaningful analysis. Simple, rule-based text classifiers or more sophisticated classifiers trained on labeled data may be used for this cleaning function[5].

Social data analytics:

User interactions on social media platforms are analyzed via social data analysis, which entails gathering and examining social metrics including average reach, total engagements, total impressions, hot topics, and more. Social data analysis usually consists of two steps first one is gathering data generated from social networking sites like we have gathered the required data for this paper from the twitter and second one is analysis of that data.

Conclusions and Reflections

This paper concludes that various classes have been analyzed based on tweets on Russia Europe. It is also useful for analyzing peoples' perspectives on a range of subjects connected to any topic. In our case study we end up with sentimental analysis of tweets about Russia and Europe without adding Ukraine in to the parameter, also excluding any time parameter, which gave us more positive sentiment than negative. Results could have been different if we consider some specific timeline and by adding more parameters such as gas pipeline or energy crisis etc. We can offer a quick automated way to assess what people believe using hash tags. Therefore, employing Big Data tools to gather information from social networks and analyze it has replaced the use of conventional database approaches.

REFERENCES

1. Adrian, C., Abdullah, R., Atan, R., & Jusoh, Y. Y. (2018).*THEORETICAL ASPECT IN FORMULATING ASSESSMENT MODEL OF BIG DATA ANALYTICS ENVIRONMENT*. Inf. Syst, 10(26), 1-51.

2. Felt, M. (2016). Social media and the social sciences: How researchers employ Big Data analytics. *Big data & society*, *3*(1), 2053951716645828.

3. Forsberg, T., & Haukkala, H. (2016). *The European Union and Russia. Bloomsbury Publishing.*

4. Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014, January). Word cloud explorer: Text analytics based on word clouds. In *2014 47th Hawaii international conference on system sciences* (pp. 1833-1842). IEEE.

5. Fan, W., & Gordon, M. D. (2014). The power of social media analytics. *Communications of the ACM*, *57*(6), 74-81.

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