Twitter Sentiment Analysis of Ukraine & Russia War

Authors: Vraj Mehalana, Vishav Patel, SanthiVijai, Nikesh Beemola
Instructor: Jongwook Woo
Department of Information Systems, California State University
Los Angeles
Tel. 323-343-2916, Fax. 323-343--5209
e-mail : jwoo5@caltstatela.edu

Abstract: Intensity of data matters when it comes to social media monitoring. Sentiment analysis is a sort of magnifier that is added to the existing data, thus converting them into context and categorizing people's emotion by type and intensity. The war between Ukraine and Russia has been a remarkably hot topic on Twitter whereby people have been expressing their thoughts and feelings. With the help of sentiment analysis, vague assumptions are eliminated by being able to explore our data more in depth and thus having a much more precise understanding of the issue being discussed. In this context, sentiment analysis helped us to understand the people's opinion in regard to their tweets as a result of them being either positive, negative, or neutral about the war.

1. Introduction

Microblogging websites have evolved to become a source of varied kinds of information. This is due to the nature of microblogs on which people post real-time messages about their opinions on a variety of topics, discuss current issues, complain, and express their sentiments about the happenings around them. Twitter is gaining popularity as a microblogging and social networking service to discuss various social issues. The purpose of the study is to explore and analyze the public sentiments related to the Ukraine war across the Twitter messages and the impact tweets make across digital social circles.

Sentiment analysis is a relatively new area that deals with extracting user opinion automatically. An example of positive sentiment is, "Courage strength resolve With Zelensky" alternatively, negative sentiment is "This is horrifying Stop Putin". Objective(neutral) texts are deemed not to be expressing any sentiment, such as news headlines, for example, "Five weeks have passed since the beginning of the Russian invasion of Ukraine". Opinions expressed in social media can be classified to determine the orientation (negative, positive, and neutral) of the posted text. Sentiment strength and intensity of the post are determined with the aim to identify the opinion and emotion of the user. With more and more people moving to the internet, huge amounts of data are being produced every second, and the challenge is to store this large data and process it efficiently in real-time to infer knowledge from this data. This paper presents different approaches for real-time and scalable ways of performing sentiment analysis using Hadoop in a time-efficient manner.

2. Related Work

In one of the works, they concentrated less on correctness and more on speed of analysis in one of the works. Splitting the various modules of data into steps and using Hadoop for mapping the sentiment analysis of big data is achieved. Opennlp was used to tag a part of speech. This labeling serves a multitude of functions. Like stop words removal: Stop words such as a, an, and this that aren't relevant in sentiment analysis are deleted in this phase. In Opennlp, stop words are labeled as _DT. This tag excludes all words with this tag. Unstructured to structured: The majority of Twitter comments are unstructured. 'gud' means 'good,' and happyyy is happy. Unstructured data records are dynamically converted to structured data records, and vowels are added. Emoticons: These are the most expressive methods for expressing one's thoughts. At this point, the emoticons' symbolic representation is transformed to words: joyful. Root form: To avoid unnecessary storage of the derived word's sentiment, the given words in a tweet are converted to their root form. The root form dictionary is utilized to accomplish this, and it is made local because it is frequently used. This reduces access time and improves the system's overall efficiency. Further the sentiment Directory is built using standard data from sentimwordnet and all conceivable uses of a single word. For example, the word "good" can be used in a variety of ways, each with its own sentiment value. So, from all of its usage, an overall emotion of good is obtained and saved in a directory that should be local to the application (i.e. in primary memory) so that time is not wasted searching for words in secondary memory storage.

In other work the sentimental Analysis is concerned with obtaining people's true feelings regarding specific products, services, organizations, films, news, events, and concerns, as well as their characteristics. Natural Language Processing, Machine Learning, Text Mining, and Information Theory and Coding are all aspects of sentiment analysis. They may categorize their unstructured data, which may be in the form of news articles, blogs, tweets, movie reviews, product reviews, and so on, into positive, negative, or neutral sentiment utilizing approaches, methods, techniques, and models of specific branches. There are three degrees of sentiment analysis. Document level - Sentence level - Aspect or Entity level Sentiment analysis at the document level is performed to determine whether the document expresses positive or negative sentiment. Fine-grained analysis is performed by Entity or Aspect Level Sentiment Analysis. The purpose of entity or aspect level sentiment analysis is to find sentiment on entities and/or their aspects. Consider the following statement: "My HTC Wildfire S phone has good picture quality but poor phone memory storage," which indicates positive emotion toward HTC's camera and display quality but negative feeling toward its phone memory storage.

In one of the works, they have carried out the following steps content Retrieval: The java Twitter streaming API is used to capture a significant volume of data. Storage: This information is saved in a specific format (HDFS: Hadoop Distributed File System) in order to create a key-value pair that is fed to the mapper in the map-reduce programming style. Hadoop Distributed File System is used to store the data. Data Processing: Data is processed over time using java and the Apache Hadoop distributed processing software architecture, as well as the map reduce programming model and the Apache Hive framework. Data Analysis: The output from the reducer phase is examined. Data Representation: Pie charts are a visual representation of categorized data. They will obtain the results in the form of Positive, Negative, and Neutral tweets in the end.

Where as in our work we directly downloaded the dataset and did the data engineering like cleaning through tableau prep. Uploaded the data on hadoop via linux further generating the output files using various queries and lastely we did the visualization using tableau and we created several insightful visualizaton.

3. Background

Twitter allows users to post real-time messages, called tweets. Tweets are short messages, restricted to 140 characters in length. Due to the nature of this microblogging service (quick and short messages), people use acronyms, make spelling mistakes, and use emoticons and other characters that express special meanings. Following is a brief terminology associated with tweets. Emoticons**:** These are facial expressions pictorially represented using punctuation and letters; they express the user’s mood. Target**:** Users of Twitter use the “@” symbol to refer to other users on the microblog. Referring to other users in this manner automatically alerts them. Hashtags**:** Users usually use hashtags to mark topics. This is primarily done to increase the visibility of their tweets.

4. Implementation and Work

Our platform specification

Cluster version	Amazon web services (EMR)
Number of Nodes	3
Memory Size	30 GB
CPU Speed	2.5 GHz
HDFS Capacity	1 TB

We acquired 20 GB of manually annotated Twitter data (tweets) from a commercial source. They have made part of their data publicly available. For information on how to obtain the data, see the Acknowledgments section at the end of the paper. They collected the data by archiving the real-time stream. No language, location, or any other kind of restriction was made during the streaming process. In fact, their collection consists of tweets in foreign languages. We eliminated the tweets with non-English and invalid locations for experiments. This leaves us with an unbalanced sample of 2GB tweets. We presented a comprehensive set of experiments for both these tasks on manually annotated data that is a random sample of a stream of tweets.

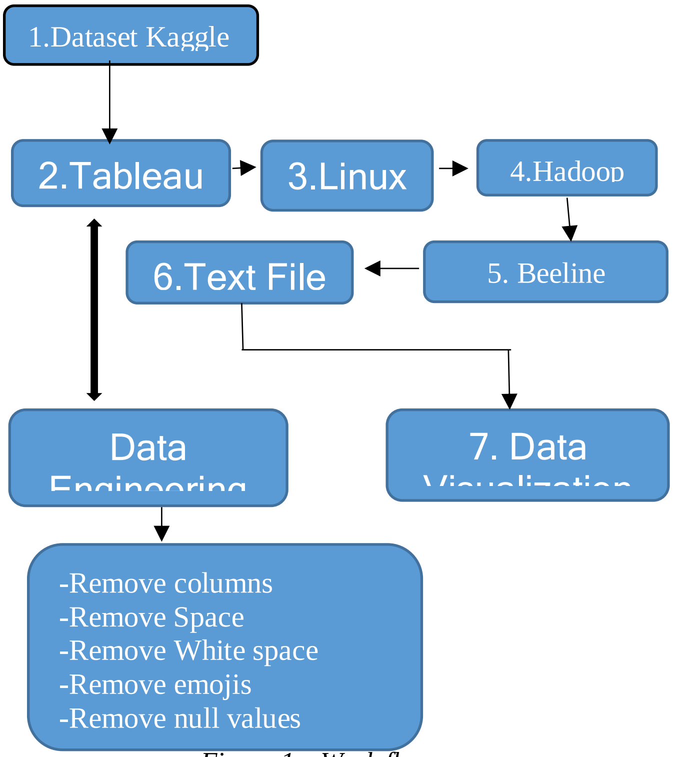


Figure 1 – Work flow

1. We initially downloaded our dataset Ukraine Conflict Twitter Dataset from Kaggle → ~20GB,
2. We proceeded to perform data engineering by utilizing Tableau Prep as our tool in order to perform data cleaning,

Tableau Prep Flow

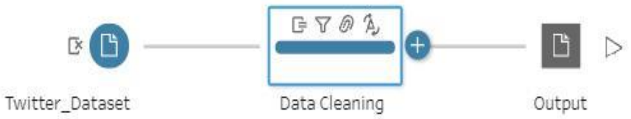
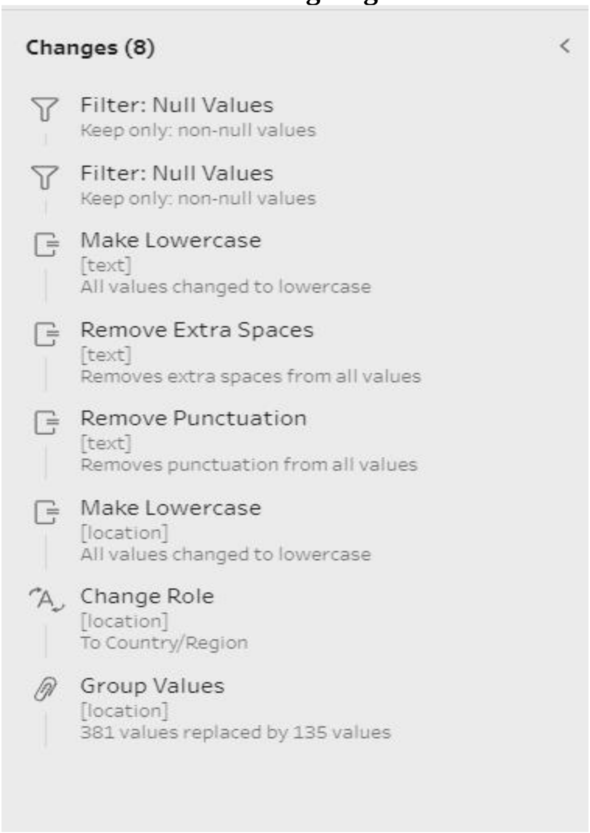


Figure 2 – Tableau prep flow

3. Right after cleaning, we uploaded the data Linux operating system - Amazon Web Service EMR,
4. Once in EMR, the data was then stored and processed into hadoop where tables were created,
5. Beeline is an interface of hive where reading, analyzing, and exporting of data takes place,
6. In the end, our data is exported,
7. Ready for visualization.

Data cleaning stages



Changes (8)

Filter: Null Values
Keep only: non-null values

Filter: Null Values
Keep only: non-null values

Make Lowercase
[text]
All values changed to lowercase

Remove Extra Spaces
[text]
Removes extra spaces from all values

Remove Punctuation
[text]
Removes punctuation from all values

Make Lowercase
[location]
All values changed to lowercase

Change Role
[location]
To Country/Region

Group Values
[location]
381 values replaced by 135 values

Figure 3 – Data cleaning stages

5. Visualization

We have used Tableau for our visualization.

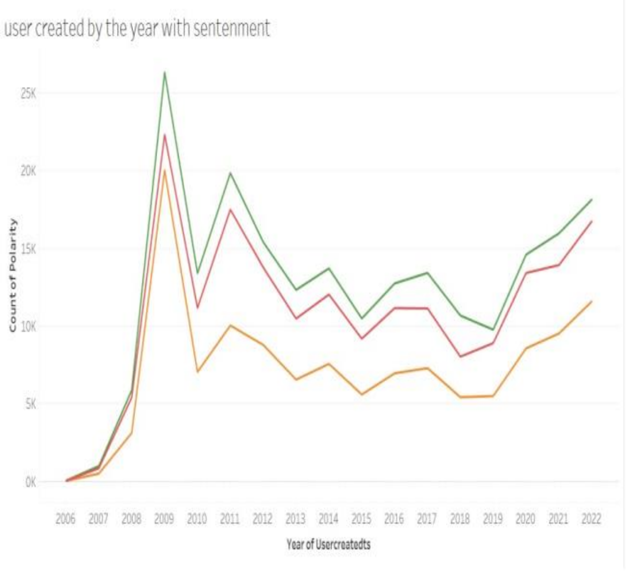


Figure 4 – User created in years

We can see the number of accounts created on initial stage and number of accounts created during 2022. If there were more fake accounts created to participate in the war, then there would have been a peak near 2022. But in this case, it's not. So there might be less number of fake accounts created during the time of war.

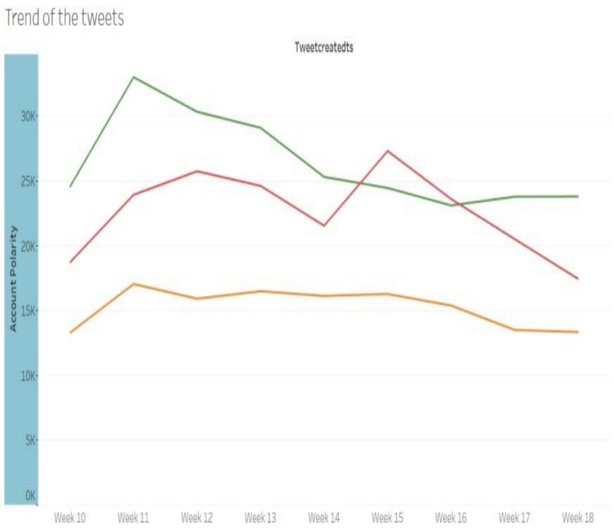


Figure 5 -Trend of tweets

We can see the number of tweets since the start of 2022 and as it gradually increases in the time of war.

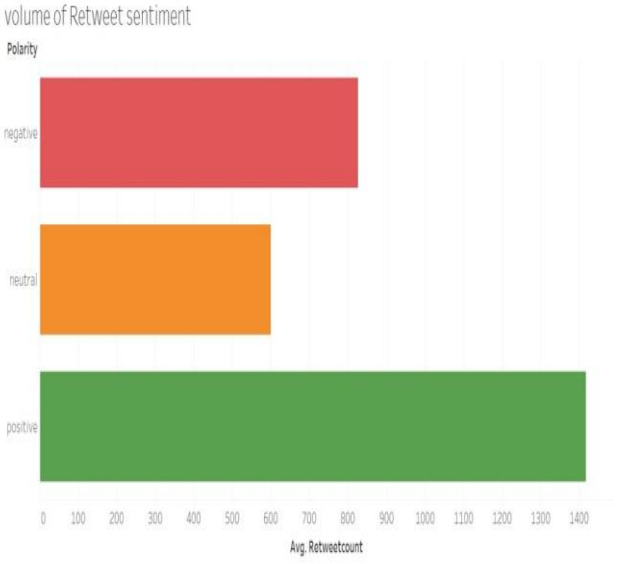


Figure 6 – Volume of Retweets

We can see the number of retweets. So from this we can conclude that how people are actively engaging in the war opinion

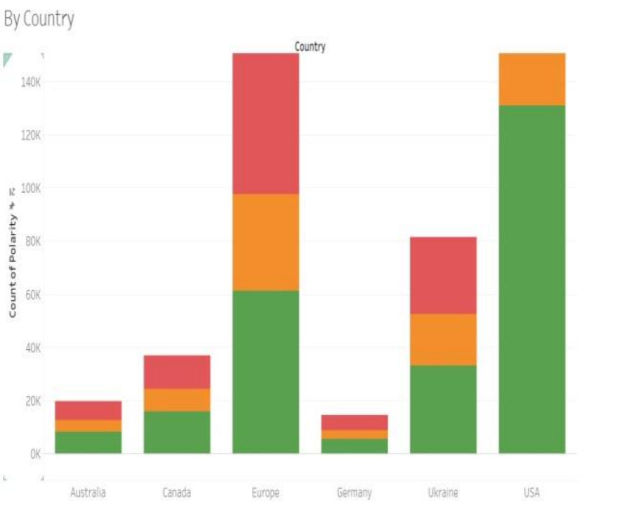


Figure 7 – Count of polarity by countries

In this graph we can see various countries and the number of tweets done from those countries. So from this we can see that people of which country are more participating and what side do they take positive, negative or neutral.

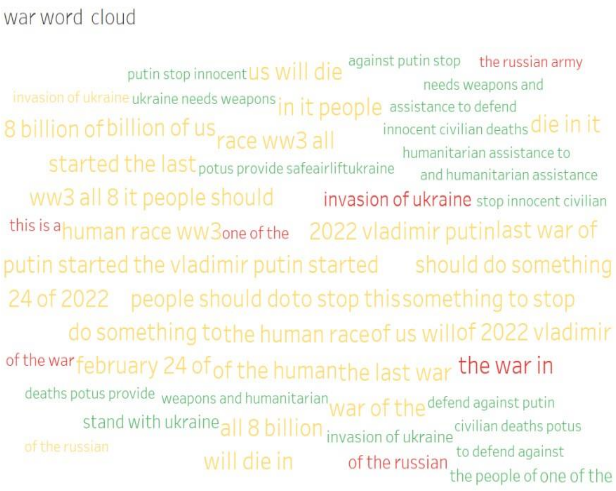


Figure 8 – Word cloud

In this we can see the words which are used lots of time and overall result of the tweets

6. Conclusion

To conclude we were able to come up with several activities of Ukraine-Russia war like which countries are supporting whom. How many people are engaged in it. What are the main influences.
This work can be carried further by using the updated dataset as it increases the number of tweets.
For more information and codes you can visit our GitHub.

References

[1] Sunil B. Mane, Yashwant Sawant, Saif Kazi, Vaibhav Shinde, "Real Time Sentiment Analysis of Twitter Data Using Hadoop"
https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.661.381&rep=rep1&type=pdf
[2] Kishor Kumar Gajula, Dr.R.Kamalakar, "An Overview of Sentiment Analysis in Bigdata Environment"
https://ijsrcseit.com/paper/CSEIT11835265.pdf
[3] Ajinkya Ingle, Anjali Kante, Shriya Samak, Anita Kumari, "Sentiment Analysis of Twitter Data Using Hadoop"
http://www.pnrsolution.org/Datacenter/Vol3/Issue6/18.pdf
[4] Dataset URL:
https://www.kaggle.com/datasets/bwandowando/ukraine-russian-crisis-twitter-dataset-1-2-m-rows
[5] Anagrams of HIVE
https://anagramscramble.com/anagrams-of/hive

Our GitHub:
https://github.com/vraj015/CIS-5200-Team-6