DEPARTMENT OF BIG DATA ANALYTICS.

ST. JOSEPH'S COLLEGE (AUTONOMOUS),

BENGALURU – 560027

ADVANCED STATISTICS PROJECT REPORT

THE WAY THE WORLD TALKS ABOUT THE RUSSIA-UKRAINE WAR

(SENTIMENT ANALYSIS ON UKRAINE-RUSSIA WAR USING R)

SUBMITTED BY:

Sarvesh S : 21BDA05
Yana Kaveramma AA : 21BDA32
Murrel Miranda : 21BDA36

YEAR: 2021-2022

TABLE OF CONTENTS

ABSTRACT

The war between Ukraine and Russia started on 23rd February 2022, it shocked the entire world and became the main topic of conversation on social networks. In this work, we use the tweets related to the war in Ukraine as data to conduct a sentimental analysis of people worldwide. The main objective of this project is to understand the psychology and behavior of societies to assist the diminishing the impact of this economic and social crisis.

1. INTRODUCTION

We tackled a very recent event. One that, as of the writing of this section, is still happening. The conflict between Ukraine and Russia has become the main topic of conversation around the world in the last few weeks. The atrocities being committed are being talked about on all social media and communication channels. All this information can be used to further our understanding of the general sentiment and find interesting insights into the way people and nations are reacting to the war. We choose to focus specifically on one social network, Twitter. We found it especially interesting as it is one of the social networks that facilitate the posting of opinions via simple texts. Tweets tend to be short, so they are easy to process. The data is easily accessible. In this way, we'll be able to perform a sentiment analysis on a massive number of tweets from worldwide. These tweets contain a great variety of opinions and could yield exciting results. The problem could be defined as "As war erupts between Ukraine and Russia, we must label and evaluate these tweets to form an overall view of the world's reaction to this event." This analysis and its results could lead to a better understanding of how people reacted to this event and can produce new opportunities.

Sentiment analysis (or opinion mining) is defined as the task of finding the opinions of authors about specific entities. In this case, we aim to identify how people feel through their tweets regarding the Ukraine-Russia conflict. The rest of this paper is organized as follows. In section II, we present the dataset to be used, data cleaning and preprocessing were performed on the same dataset. In Section III, we talk about the sentiment analysis and show the results. Finally, in Section IV, we have the conclusions for this paper.

2. DATA

1. DATA COLLECTION

Despite the fact that there were lots of speculations about the probable invasion in the press, it came as a complete shock for many other Russians.

Kaggle dataset named “RUSSIA-UKRAINE WAR TWEETS DATASET (65 DAYS)” was used in this project. It was created by a Russian named Daria Purtova as he wanted to see the evolution of the discussion around Ukraine and Russia. So he parsed quite a big number of tweets and hence this dataset was created. Despite the fact that there were lots of speculations about the probable invasion in the press, it came as a complete shock for many other Russians.

How is the dataset created?

> Max 5000 for the day
> Words used for the search ['ukraine war', 'ukraine troops', 'ukraine border', 'ukraine NATO', 'StandwithUkraine', 'russian troops', 'russian border ukraine', 'russia invade']
> For each search, a separate CSV was created
> Dates: from 2022 - 01 - 01 to 2022 - 03 -06

Dataset consists of 9 separate CSV files for each search mentioned above.

2. DATA CLEANING

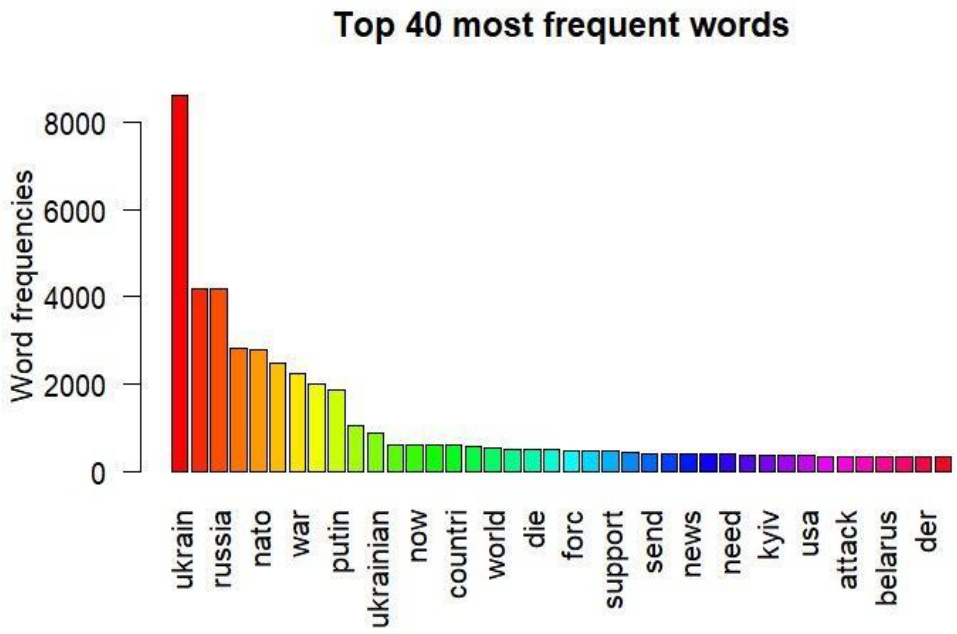
We began the cleaning by removing the columns that we did not want. The entire dataset had 29 columns. We dropped the columns which are not necessary. The example containing NaN values was also eliminated from the dataset. these values were found in the ‘location’. At last, we got 3 lakh rows and 18 columns left.

3. DATA PREPROCESSING

Due to the restriction of corpus we again sampled our data into 11,000 rows, the corpus is nothing but a collection of documents. We then did tokenization and later we incorporated that with the term-document matrix.

Some of the insights from our data preprocessing are:

1. The number of tweets for the top 40 most frequent words in the dataset



2. Association table showing the co-occurrence of words in multiple documents.

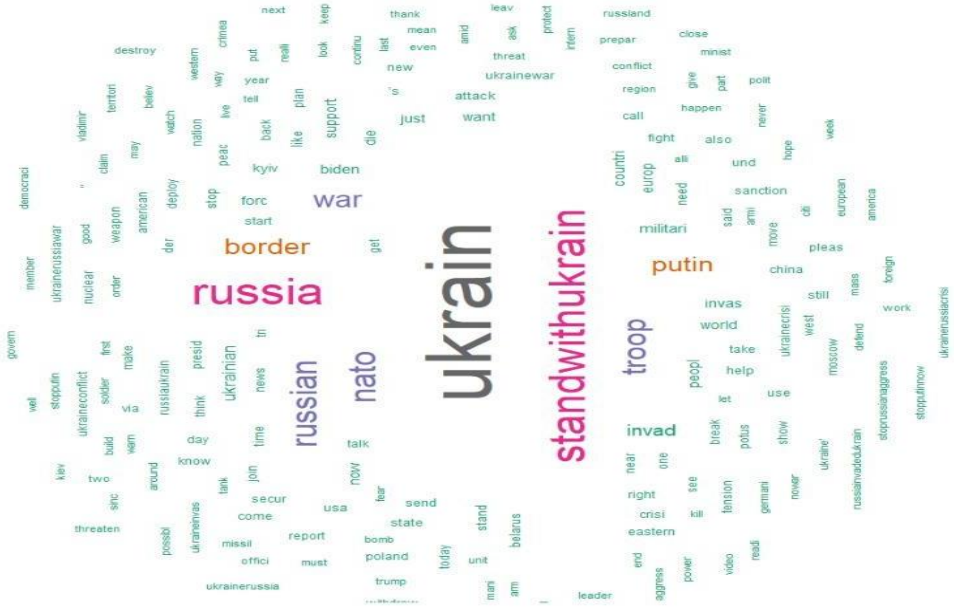
```
> findAssocs(TextDoc_dtm, terms = c("standwithukrain","russia","border"), corlimit = 0.15)

$standwithukrain
numeric(0)

$russia
ukrain  invad   nato
  0.35   0.34   0.23

$border
   cross  russian   ukrain    along    troop     near    amass southern    guard     mass
    0.24     0.24     0.21     0.20     0.20     0.20     0.18     0.17     0.16     0.15
```

3. Word Cloud showing the top 200 used words in the tweets.

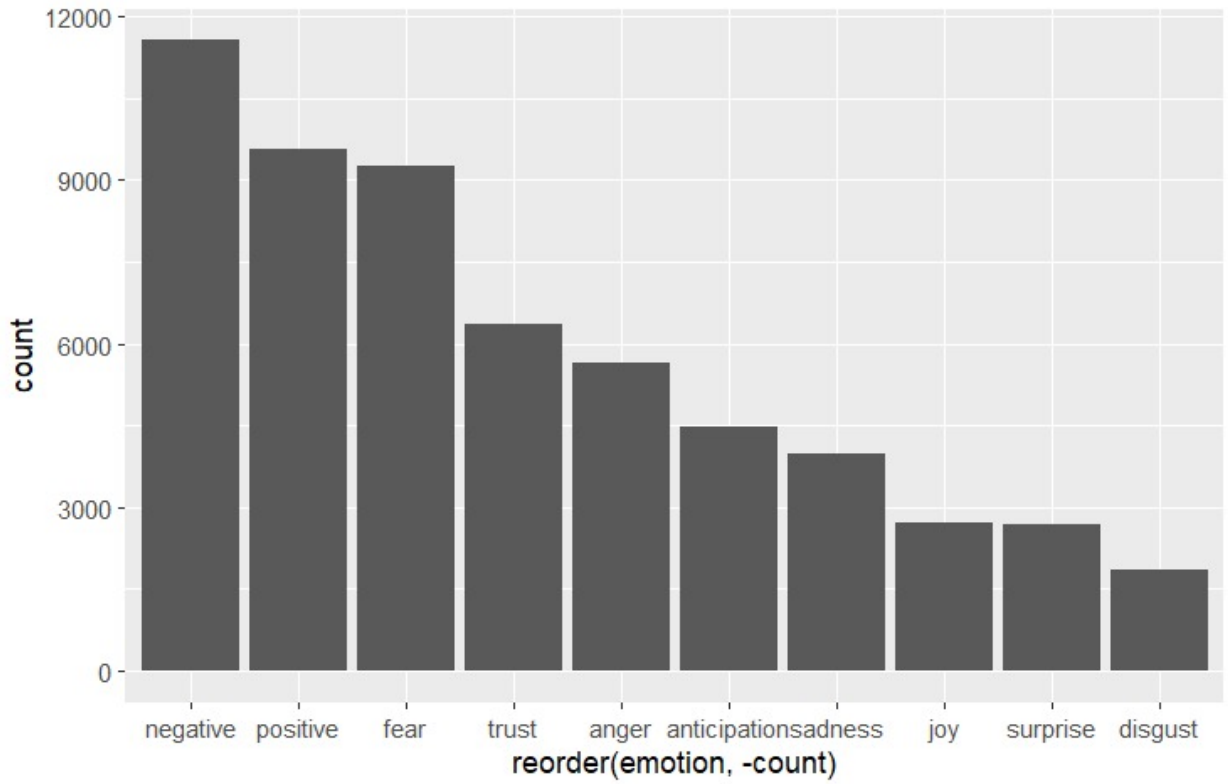


In a word cloud, the larger the size of the word the higher the number of times it's been used. Hence in the above word cloud we see that the word "Ukraine" is most used in the tweets followed by "stand with Ukraine" and "Russia".

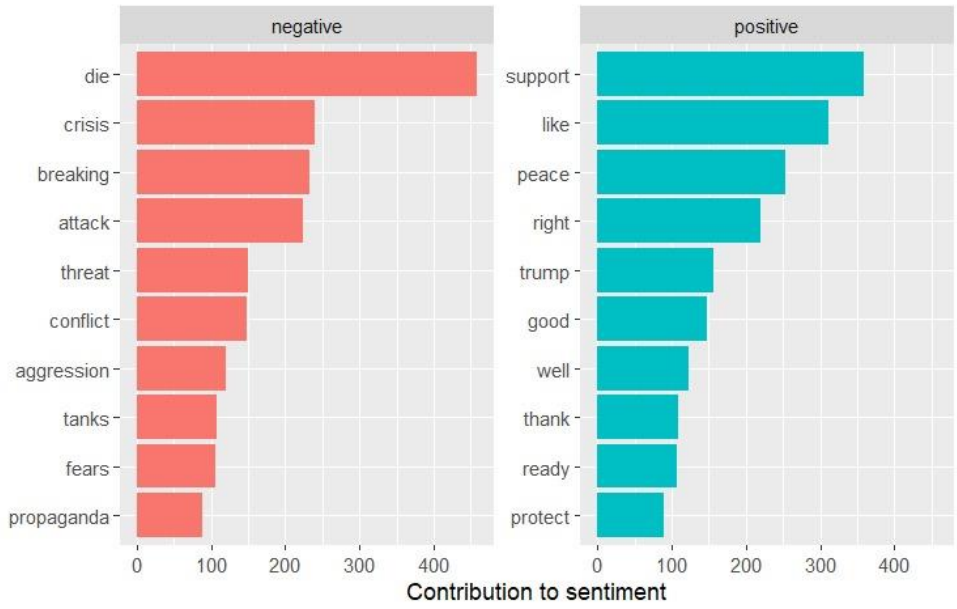
3. EMOTION CLASSIFICATION

Sentiment analysis (or opinion mining) is defined as the task of finding the opinions of authors about specific entities.

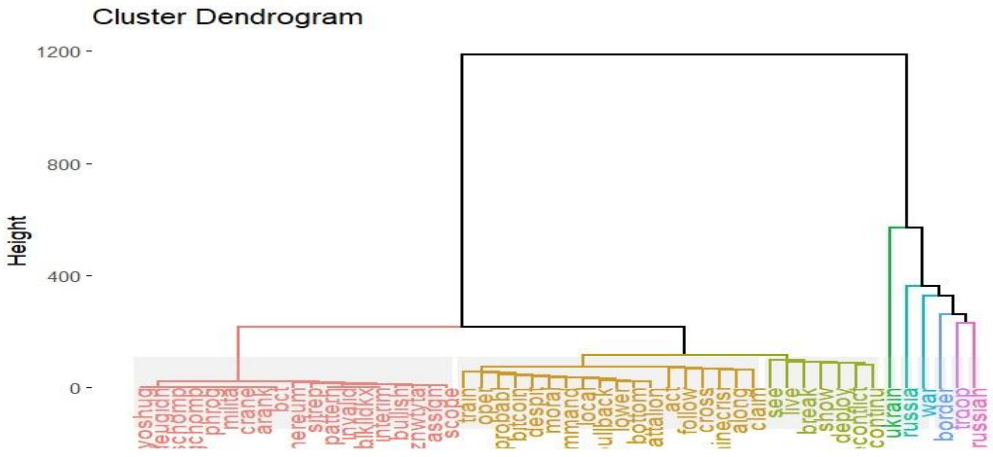
In this case, we aim to identify how people feel through their tweets regarding the Ukraine-Russia conflict. The entire document is associated with 8 emotions (anger, fear, sadness, etc.) and negative and positive.



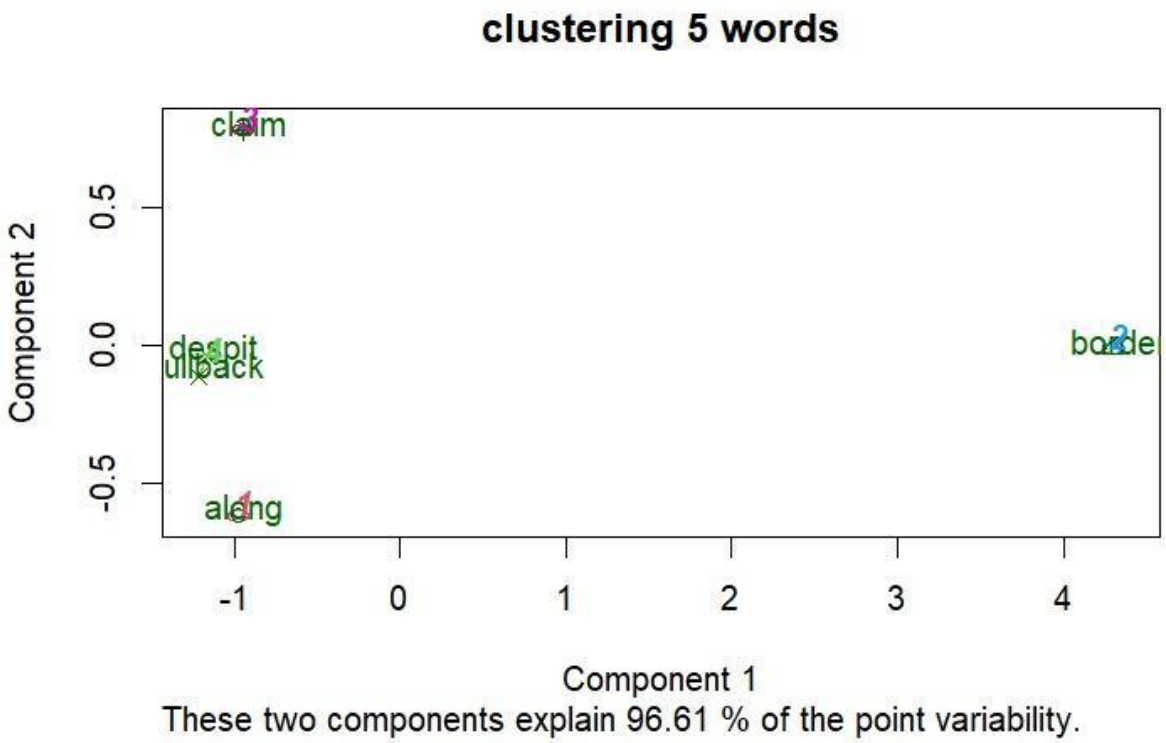
The above image show that the highest amount of tweets present was ‘negative’.



The above image show that the highest number of words used in negative tweets is "die" and the highest number of words used in positive tweets is "support".



Dendrogram is a representation of hierarchical clustering and lower the distance between the values the cluster will form cluster and higher the distance between them the cluster will form at last for distance we use Euclidian it follows Pythagoras theorem (ac2 =Ab2+bc2) to find the distance and the formula for calculating Euclidian distance is sqrt of (x2-x1)^2 -(y2-y1)^2



Here we took 5 words which had maximum frequency to cluster, at first, we need to define how many clusters we want to make and here we have chosen 4 clusters. Then we have to choose a random point and we calculate distance of that point and the values from that point. We then have to cluster with respect to minimum distance from that point and after this again we need to calculate the mean and find minimum distance. Based on that we need to club and do iteration till the migration stops.

4. CONCLUSION

In this work, we performed sentiment analysis of the tweets related to the Ukraine-Russia war from all around the world. Once the tweets were cleaned, we preprocessed the data to have a better understanding of the presented data. Judging by the results of the population data as seen on the media we believe we obtained great results as our sample data results show similar results.

The sentiment analysis helps us understand that, as predicted, a high percentage of the tweets related to the Ukraine-Russia war were negative. Even though the current circumstances are horrible, people around the world, one way or another, tried to contribute by spreading some positiveness in their tweets. Judging by the results of the population data as seen on the media we believe we obtained great results as our sample data results show similar results.