SENTIMENT ANALYSIS OF RUSSIA AND UKRAINE CONFLICT

Ashokkumar Narayanappa
Department of Computer Science
University of Exeter
Exeter, UK
an528@exeter.ac.uk

Riccardo Di Clemente
Department of Computer Science
University of Exeter
Exeter, UK
r.di-clemente@exeter.ac.uk

Massimo Stella
Department of Computer Science
University of Exeter
Exeter, UK
m.stella@exeter.ac.uk

***Abstract*— Text analysis or opinion mining is one of the computational study of people's opinions. This paper shows the different opinion of the people across the globe for crisis between Russian and Ukraine war. Here we are considering twitter dataset to analyze the opinion of people. Based on the text from users we will decide whether they are positive towards war or negative or they are neutral. Sentiment analysis is one of the most effective kinds of research areas in NLP and text mining. We can see the mind set of people and their opinion is against Russia or Ukraine or any other country. Here people are stating their opinion in different languages but we will consider only English tweets.**

AIM:

The following points are main objective of this project:

- Analysis of people opinions between Russia invade Ukraine.
- Collect twitter dataset for particular time period.
- Discussion about Russia and ukraine crisis.
- Represent people tweets in three categories (positive, negative or neutral)
- Considering only English language tweets.
- Sentiment analysis of opinions.
- Building a model and calculate accuracy of the text analysis.

I. INTRODUCTION

Russia and Ukraine war is become one of the major issues around the world, because it affecting economy of countries and they are facing high price in assets (petrol, natural gas, oil etc.). People from across the globe they are stating their opinion in social media, here we will consider only twitter dataset because twitter is identified as one of the major platforms for social networking service. Here we are going to apply text stemming, lemmatization and sentiment analyzer for collected dataset. Once after the sentiment analysis we will display some diagrammatic representation of text and will see what are all the common words in positive side and negative side.

The Russia and Ukraine conflict is still ongoing, As a result of the conflict, social media has become an important tool in sharing people's opinions. Studies of prior large-scale information discourses during crises and social movements have commonly used online social networks (OSN) as a source of data, especially with regards to 'information war', where the platforms can serve as sources of propaganda and misinformation. Our open dataset on this ongoing crisis will facilitate timely analysis of the situation. Using this data, a study of the political discourse, opinion mining, and the spread of misinformation on Twitter can be accomplished.

II. DATASET

In this analysis we are considering last 3 months of tweets starting from 1st January 2022 to 6th march 2022. We are not taking all tweets from the source, we used some keywords to segregate the tweets and collected only those tweets for analysis. We are have collected more than a million of tweets regarding Russia and Ukraine crisis. Any tweet containing any of the requested words is returned via Twitter's streaming API.

KEYWORDS: 'ukraine war', 'ukraine troops', 'ukraine border', 'ukraine NATO', 'StandwithUkraine', 'russian troops', 'russian border ukraine', 'russia invade'

III. DATASET SUMMARY AND EXPLORATION

Over more than1.2 million tweets were collected by the 6th of March using the keyword phrases mentioned above. During the data collection process, no language filters or geo-filters are applied. Thus, tweets in several languages from different regions are included in the dataset. We can see in below plot, majority of tweets are in English language, So we will continue with English tweets and we omitted other languages tweets.

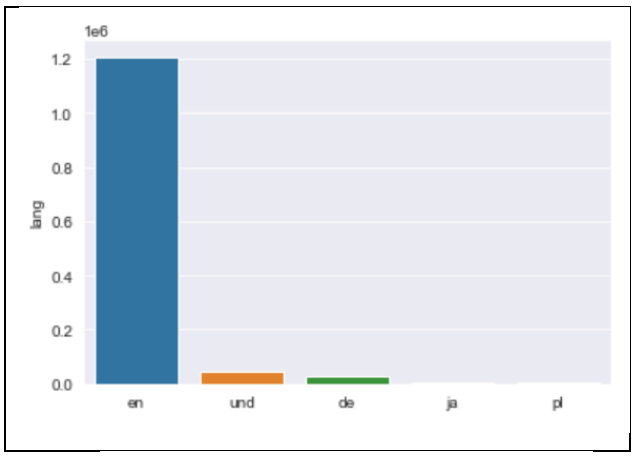

Fig 1. A bar plot with languages used in the dataset

The dataset contains lot of columns, For Data pre-processing step I am dropping NULL and duplicate values from 'content' column. We don't need all the columns in the data so we extracted only content column and created new data frame. Before starting text processing, it is better to convert all the texts into lowercase letters. We followed below steps in data pre-processing.

- Convert all 'content' field data into lower case letters.
- Apply Tokenization
- Remove punctuation from the text.

- Remove the stopwords from the dataset.
- Apply Stemming process to the dataset.
- Apply Lemmatization

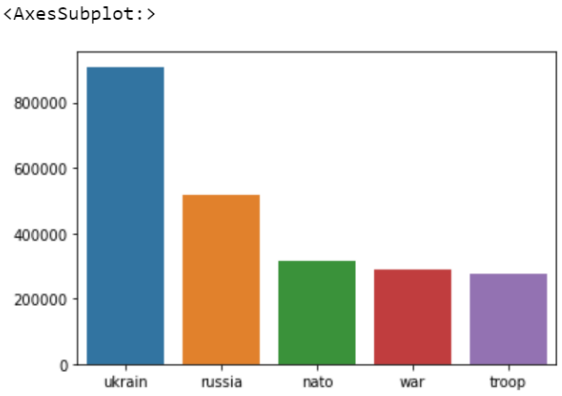

Fig 2. A bar plot for frequent words

The above plot shows the frequent words in a data set. As we can see in the data Ukraine is most used word in the dataset followed by **Russia** and **nato**.

The below picture shows the word cloud after data pre-processing.

Fig 3. A word cloud of tweets – After data cleaning

IV. SENTIMENT ANALYSIS

After data cleaning and pre-processing, we are going to do sentiment analysis of dataset. Here we are calculating sentiment scores to analyse whether the tweets are positive, negative or neutral. We used SentimentIntensityAnalyzer() to get the polarity score of each tweet. Based on the polarity score we decide whether the tweet is towards positive or negative or neutral. The below code snippet shows the calculating of polarity scores of positive, negative and neutral.

```
nltk.download('vader_lexicon')
sentiments = SentimentIntensityAnalyzer()
df["Positive"] = [sentiments.polarity_scores(i)["pos"] for i in df["text_lemmatized"]]
df["Negative"] = [sentiments.polarity_scores(i)["neg"] for i in df["text_lemmatized"]]
df["Neutral"] = [sentiments.polarity_scores(i)["neu"] for i in df["text_lemmatized"]]
```

Snippet 1. Shows how to calculate Positive, Negative and Neutral Scores.

After getting scores of sentiment analysis, we come to know most of the twitter users are against the war but some people are showing interest towards the attack and some people are neutral. Neutral means either they are not against the war are towards the war. The below plot shows the positive, negative and neutral states.

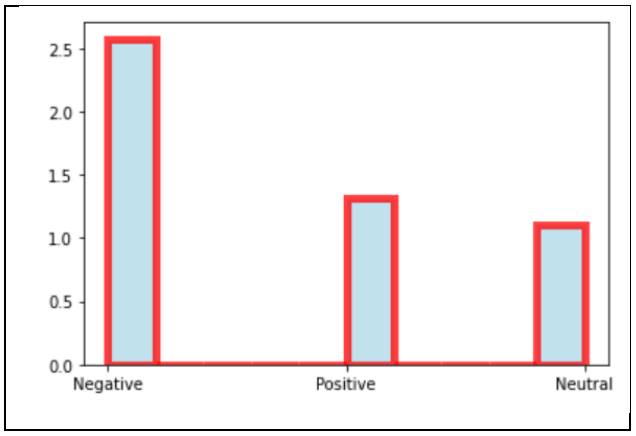

Fig 4. A Plot shows the different opinions of twitter user regarding Russia and Ukraine war

From the above plot we can say that most of the twitter users are against the war or attack, but some people showing positive opinion about war and some people are neutral in their opinion. Majority of the user's data shows Negative opinion towards the war.

V. WORD OF IDENTIFICATION

Here we will see the most occurrence of words or keywords in positive, negative and neutral tweets.

Positive	Negative	Neutral
Russian	Poland	May be
Border	Troop	India
Soviet	Zelensky	Eastern
Moscow	Ukrainian	Joe biden
Nato	Russian	Ukrainrussian
Force	Pledg	America
Military	Black sea	Possible
vladimir	Pull back	Russian

Table 1: Keywords for Positive, Negative and Neutral words

We listed out 8 keywords from all three categories from the sentiment analysis. As we can see from the above tables **Russian** keyword is appearing in all opinions.

VI. MODEL ANALYSIS USING SVM & LOGISTIC REGRESSION

Using SVMs for classification and regression, support vector machines are simple supervised machine algorithms. In essence, SVM determines the boundary between data classes by finding a hyperplane that separates them. Here we will find accuracy score and F1 score based on the SVM model generated vectors.

	Precision	Recall	F1-score	Support
Negative	0.86	0.86	0.86	3516
Positive	0.86	0.91	0.89	1399
Neutral	0.78	0.74	0.76	1085
Accuracy	0.81	0.82	0.84	6000
Micro avg	0.83	0.84	0.84	6000
Weighted avg	0.83	0.84	0.84	6000

Table 2: Classification report for twitter data set

From table 2, we can see the classification report of people opinion data from twitter. As per the table we can see there are more opinion are towards Negative and less neutral and positive stands in the second position.

Now we will find prediction of the trained dataset and will calculate accuracy of the predicted model. I am just showing the small code snippet for prediction and accuracy of the SVM model.

```
: xTrain_tweets_tVect = t_tweet.transform(X_train)
tweet_Model = svm.SVC(kernel='linear')
tweet_Model.fit(xTrain_tweets_tVect, y_train)
prediction_tweets = tweet_Model.predict(t_tweet.transform(X_test))
print('accuracy :', accuracy_score(y_test, prediction_tweets))

accuracy : 0.8418333333333333
```

Snippet 2: SVM model to calculate accuracy.

From the above snippet code, we can see accuracy score is 84%. As a good model the accuracy must be more than 70%, so we got good score in SVM model.

Lets see how confusion matrix will give results.

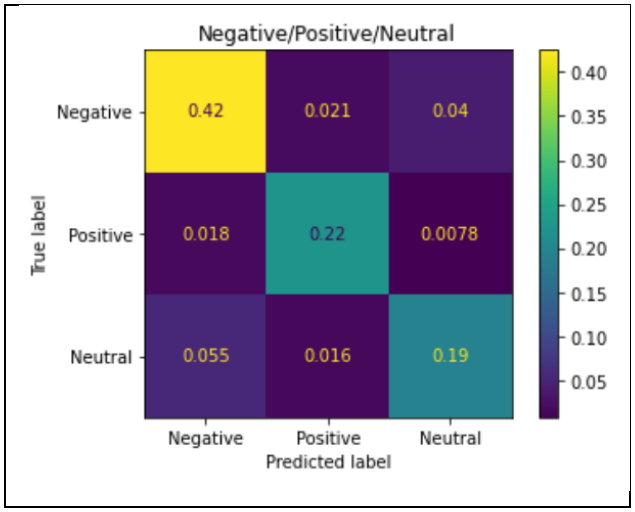

Fig 5: Confusion matrix

The above figure (confusion matrix) also says that the number of negative opinions towards the war is more than positive and neutral.

Now let's apply logistic regression model to twitter data set and will discuss the results.

```
accuracy = test_logistic_regression(test_x, test_y, freqs, theta
print('The accuracy of Logistic Regression is :', accuracy)

The accuracy of Logistic Regression is : 0.8210755813953488
```

Snippet 3: Logistic Regression accuracy calculation

In Snippet 3 we calculated the accuracy of dataset using logistic regression. Here the accuracy is slightly varied compared to SVM model. But accuracy score is acceptable and we can conclude it's a good model.

Now we will just see an example how prediction works in sentences.

```
def predict_tweet(tweet, freqs, theta):
    x = extract_features(tweet, freqs)
    # make prediction
    y_pred = sigmoid(np.dot(x, theta))
    return y_pred
```

Snippet 4: predict_tweet function

The above code snippet shows the implementation of predict_tweet(), In this function we are call sub-function among that sigmoid function will play an important role to predict whether its positive or negative.

Now we are going to call this predict_tweet() by passing some predefined text.

```
txt = 'Russia should not pull back their millitary forces from ukraine border'
print(process_tweet(txt))
y_hat = predict_tweet(txt, freqs, theta)

if y_hat > 0.5:
    print('Positive')
else:
    print('Negative')

['russia', 'pull', 'back', 'millitari', 'forc', 'ukrain', 'border']
Positive
```

Snippet 5: Demonstration of predict_tweet()

As we can see the code of Snippet 5, the result is positive, because our Sentence is "Russia should not pull back their forces from Ukraine border". This sentence shows positive approach towards war.

Conclusion:

To conclude, As we can see from the above results, large number of people opinion are against the war (negative towards war). But, some of the people are justifying their opinion to invade Ukraine (positive towards war). Similarly other set of people they are neutral in their opinion, Neutral means they stay away from the conflict between Russia and Ukraine and their sentences like "Border conflict is common for every country, Russia and Ukraine ministers should talk and resolve this crisis". As a result, we can say that most of the people wants peace towards both nations, from the analysis and plots we can say around 55%-60% of the twitter user against the attack(negative), 19%-25% are supporting war (positive) and 15%-20% people are neutral.

REFERENCES

[1] https://thecleverprogrammer.com/2021/09/13/twitter-sentiment-analysis-using-python/

[2] https://www.kaggle.com/datasets/foklacu/ukraine-war-tweets-dataset-65-days

[3] https://www.kaggle.com/code/ollibolli/ukraine-conflict-twitter-eda

[4] https://medium.datadriveninvestor.com/sentiment-analysis-of-tweets-using-logistic-regression-in-python-dba10a682556

[5] https://arxiv.org/pdf/2203.02955.pdf

[6] https://www.educative.io/edpresso/how-to-use-svms-for-text-classification

NOTE: Please find the code in the below URL

https://colab.research.google.com/drive/1IyR1Zz0f_wwyjhCLBsCpjXHAzf9dUl5o?usp=sharing