A SOCIAL MEDIA SENTIMENT ANALYSIS ON THE WAR OF UKRAINE AGAINST RUSSIA

MIE 1624 Project Report

Group 10

Andrew Du	1005330991
Haowen Li	1008399289
Siyuan Wang	1007876906
Steven Xie	998979627
Yunhao Zheng	1008057326
Yunjie Xu	1008601951
Zhijian Zhu	1002386508

FALL 2022

Table of Contents

1. Background and Methodology

The Russia Ukraine War was a prolonged hybrid war between Russia and Ukraine that began since February 2014 and was carried out in a low-intensity war in the early stage. On February 24, 2022, Russian President Vladimir Putin mobilized the Russian army to invade Ukraine on the grounds of "demilitarization and de-Nazification", turning the conflict into a full-scale war and rapidly developing into the largest war in Europe since World War II. The war is widely viewed as an invasion, though some people support the justice of Putin's decision on starting the war.

Sentiment analysis can be used to analyze tone for text. In this project, sentiment analysis is used to detect trends in public opinions on Russia's Invasion to Ukraine by analyzing posts on social media. Then, based on the results, we will make recommendations to improve Ukraine's image, and gain more international support to defend their territory against Russians.

In the first part, we use the provided dataset of classified tweets to train four sentiment analysis models. The four classification algorithms chosen are logistic regression, decision tree, random forest, and Gaussian Naïve Bayes. We compare the performance of the four classification algorithms and select the best model to be used in part 2. In the second part, we use scrapped comments and posts related to the Russian Ukraine War from Twitter and Reddit and implement pretrained model VADER (Valence Aware Dictionary and sentiment Reasoner) to perform sentiment analysis. Then the best model from part 1 will be implemented on the same dataset and make comparison of the results between our best model and the VADER model. The third part of this project is to identify factors/reasons/topics that drive sentiment. Latent Dirichlet Allocation (LDA), and Word Cloud are utilized to discover the most influential factors/reasons/topics. The key results of the previous three parts will be visualized in the last part and used to explore the most concerning topics and provide suggestions to the Ukrainian government and international NGOs.

2. Development of the Sentiment Classification Model

2.1 Training the Model

The provided dataset of classified tweets was used to train our sentiment analysis model. The dataset contains binary classes, where 0 and 1 are corresponded to negative and positive sentiments, respectively.

Before training the classification model, it is necessary to implement the pre-processing and feature engineering to prepare the data for machine learning algorithms. Therefore, the HTML tags, hashtags, user tags, emojis, digits, punctuations and stop words were removed from the tweets before all words were lemmatized. After using TF-IDF, the top 500 words with the highest frequency are chosen to train the classification model.

Then, we trained four classification algorithms (Logistic regression, decision tree, random forest, and Gaussian Naive Bayes) on the prepared training dataset (containing 250,420 samples). After model tuning, the optimal results for each algorithm for training and test sets are plotted in Figure 1 and Figure 2, respectively. By comparing the metrics across the four machine learning models, we can see that all models performed well on the training and testing datasets. No overfitting or underfitting was observed. Among all the methods, Logistic Regression has a relatively better performance. In addition, Logistic Regression took the least computational cost (fastest in terms of cross-validation and model fitting) among the four algorithms. Therefore, the Logistic Regression model is selected to be the best model for sentiment analysis and will be used in the following parts of this project.

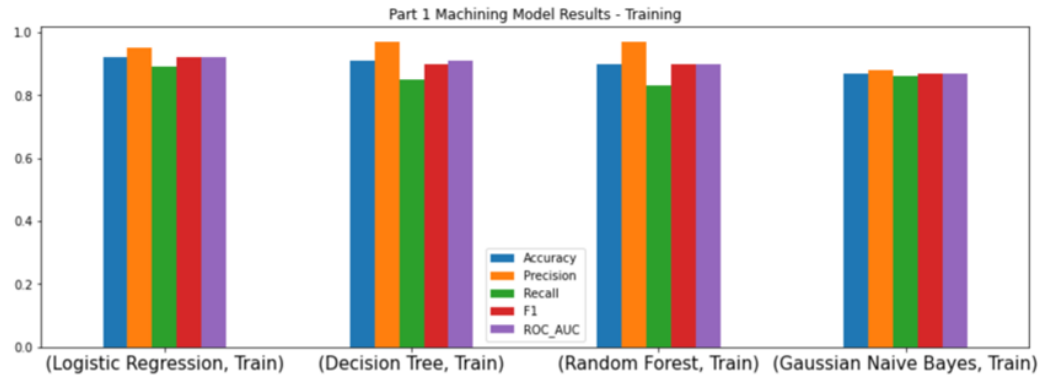

Figure 1. The Training Results of the Machine Learning Models

Figure 2. The Test Results of the Machine Learning Models

2.2 Model Application

In this section, we will apply (1) the logistic regression (LR) model from the last section, and (2) VADER (a pre-trained sentiment classification model) to postings and comments on Twitter and Reddit that are related to Russia's war in Ukraine. The web-scraped Twitter dataset contains 29,562 Tweets, and the scraped Reddit dataset contains 3,231 posts. The results from both models for each dataset will be compared and discussed.

2.2.1 Application of VADER and the Developed Logistic Regression Model

VADER (Valence Aware Dictionary and sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool. VADER relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores. After assigning sentiment scores to each word of the text, VADER calculates the compound score, which is the normalized sum of positive, negative, and neutral sentiment scores, and we used this compound score to perform classification.

2.2.1.1 Twitter Dataset

The sentiment composition of the VADER-classified Tweets is shown in Figure 3, in which the Tweet-Ternary pie chart shows that among all Tweets, 30% is negative, 33% is positive, and 37% is neutral. After ignoring the neutral tweets, we yield the Tweet-Binary pie chart, showing that the remaining 18,521 Tweets, with a definite polarity, have a sentiment composition of 47% negative and 53% positive. Then, we applied the LR model to those 18,521 Tweets, and the results are illustrated in the pie chart on the right side of Figure 4 showing a sentiment composition of 24% positive and 76% negative. For comparison purposes, the sentiment composition (or, class distribution) of the initial given Twitter dataset used for training is shown on the left side of Figure 4 showing a sentiment composition of 67% positive and 33% negative.

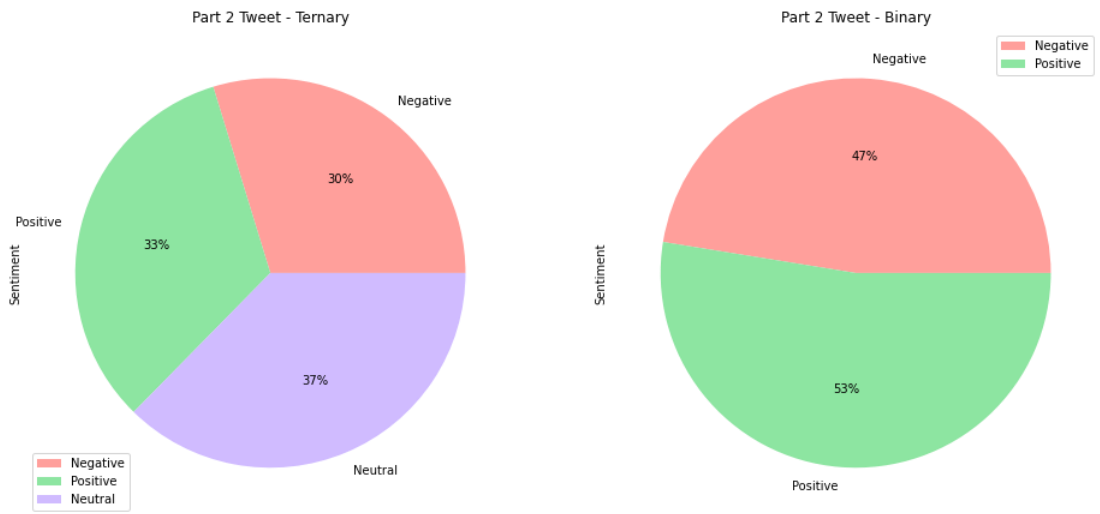

Figure 3. War-Related Twitter Sentiment Classification using VADER

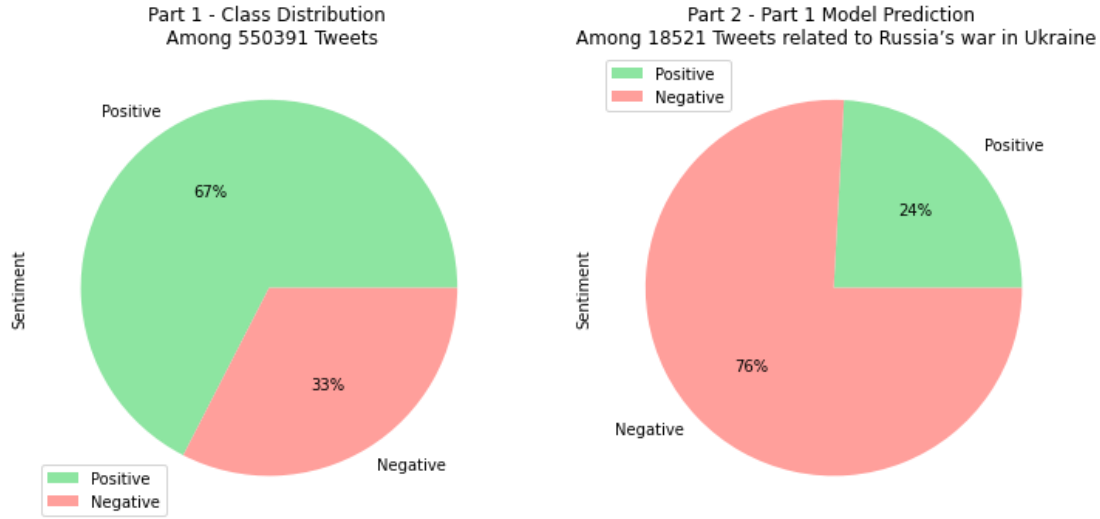

Figure 4. War-Related Twitter Sentiment Classification using the Logistic Regression Model

2.2.1.2 Reddit Dataset

The sentiment composition of the VADER-classified Reddit posts is shown in Figure 5, in which the Reddit-Ternary pie chart shows that among all Reddit posts, 43% is negative, 31% is positive, and 25% is neutral. After ignoring the neutral posts, we yield the Reddit-Binary pie chart, showing that the remaining 2,410 Reddit posts, with a definite polarity, have a sentiment composition of 58% negative and 42% positive. Then, we applied the LR model to those 2,410 Reddit posts, and the results are illustrated in the pie chart on the right side of Figure 6 showing a sentiment composition of 17% positive and 83% negative. For comparison purposes, the two pie charts in Figure 5 are included in Figure 6.

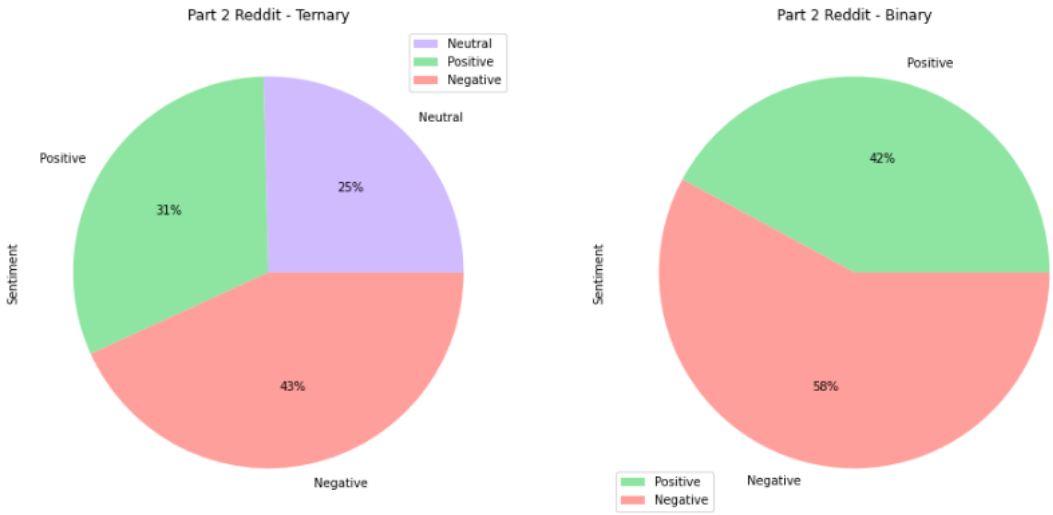

Figure 5. War-Related Reddit Sentiment Classification using VADER

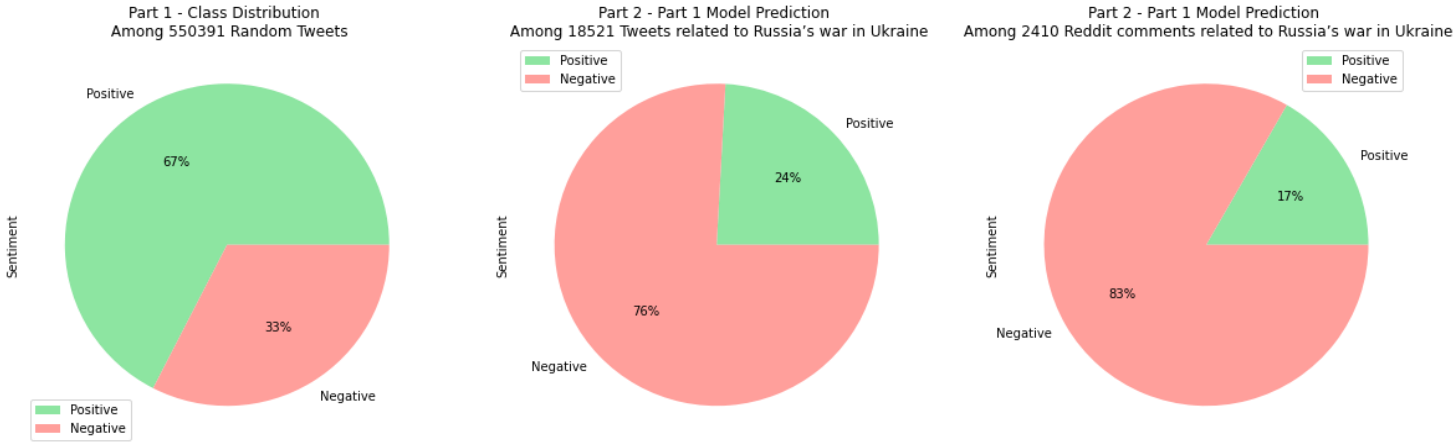

Figure 6. War-Related Reddit Sentiment Classification using the Logistic Regression Model

2.2.2 Classification Results Comparison and Discussion

The model application results are summarized in Table 1 in terms of the negative sentiment percentage of 18,521 Tweets and 2,410 Reddit Posts.

Table 1. Model Application Results – The Negative Sentiment Percentages of 18,521 Tweets and 2,410 Reddit Posts

	Twitter	**Reddit**	**Col % Difference**
VADER	47%	56%	19.1%

Logistic Regression (LR)	76%	83%	9.2%
Row % Difference	61.7%	48.2%	

By comparing the VADER row to the LR row in Table 1, we noticed that the LR model classified 61.7% and 48.2% more samples as negative sentiment than VADER for the Twitter and Reddit datasets, respectively. These discrepancies are probably because the LR model was trained on a dataset unrelated to the Russian-Ukraine War, and thus the LR model could not pick up some war-related positive lexicon. On the other hand, VADER's dictionary is more inclusive and it's intelligent enough to understand the basic context of words. By comparing the Twitter column to the Reddit column in Table 1, we can see that Reddit has higher negative sentiment percentages than Twitter regardless of what classification model was used. This similarity in trend justifies the application of the LR model even though we have discussed the classification discrepancies between this model and VADER.

Moreover, as shown in Figure 6, the negative sentiment rate of 55,0391 random Tweets is only 33%, and we can use this as a baseline for normal Twitter sentiment composition as the sample size is sufficiently large. The LR-predicted negative rate of the 18,521 war-related tweets is 76% which is more than twice as large as the baseline percentage, implying that the LR model was able to capture people's negative emotions regarding the war. Upon classifying the sentiments of the war-related social media datasets, we can conclude that most people on social media are showing negative emotions against Russia's war in Ukraine, and Reddit users are showing more negative emotions than Twitter users.

3. Factors and Topics Identification

3.1 Implementation of LDA

To identify topics and relations to polarized sentiments, the team implemented the Latent Dirichlet allocation (LDA) model on the LG-classified war-related Twitter and Reddit datasets. We also used the LDAvis library in Python to visualize the LDA results.

LDA is a probabilistic model that groups words from a set of documents into topics. The LDA model assumes that each document has a distribution of topics, and each word in the document comes from this topic distribution. It is an unsupervised model that learns by maximizing the probability of each topic given the word distributions.

3.2 Interpretation of the Identified Topics/Factors

From previous sentiment analysis, we noticed that the majority of people on social media are expressing negative feelings about the war. In this part, we want to discover what are the factors and topics that make people so frustrated about the war.

Factor 1 - Russia cuts off gas exports to Europe

This is identified as Topic 26 from Twitter's negative posts and Topic 8 from Reddit's negative posts. We can see the keywords here are pay, gas, Germany, price, Europe, etc. This group of words shadows the Russians cutting off the gas exports to Europe. This year, Russia cut its gas supplies to the EU by 88%. The Nord Stream 1 and 2 have been closed and destroyed respectively within months. A lot of people in Europe are suffering and will suffer the coldness of this upcoming winter. [1]

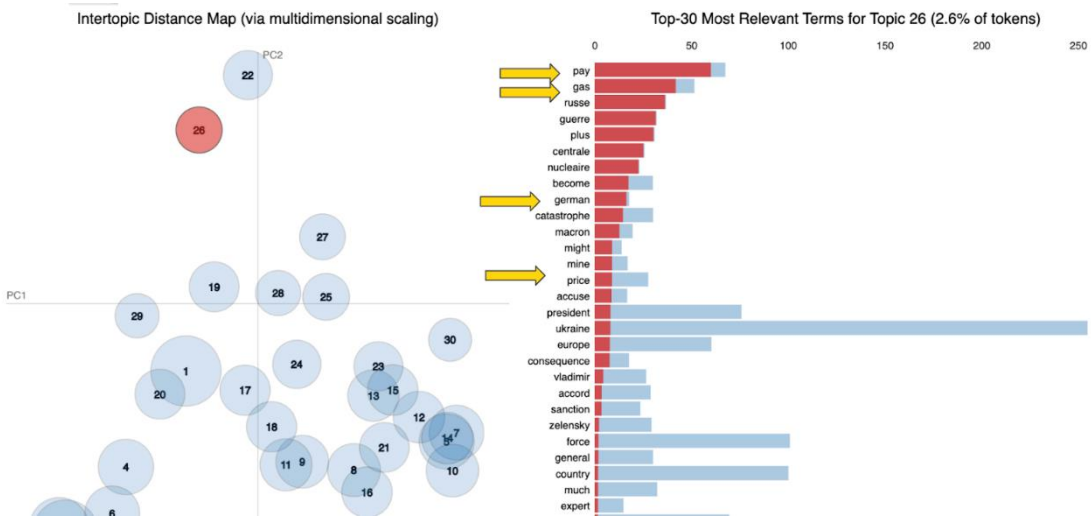

Figure 7. Twitter Negative Topic 26 - Natural Gas Supply

Factor 2 - Russia takes the control of the Zaporizhzhia nuclear plant

This is identified as Topic 1 from Twitter's negative posts and Topic 20 from Reddit's negative posts. Here we can see keywords like nuclear, plant, power, Zaporizhzhia, catastrophe, etc. It demonstrates that Russia's invasion has added instability to the world since Zaporizhzhia nuclear power plants in warring areas are being affected, which has increased the insecurity factor and raised global concerns. It is insidious of the Russian troop to seize control of the nuclear plant. [2]

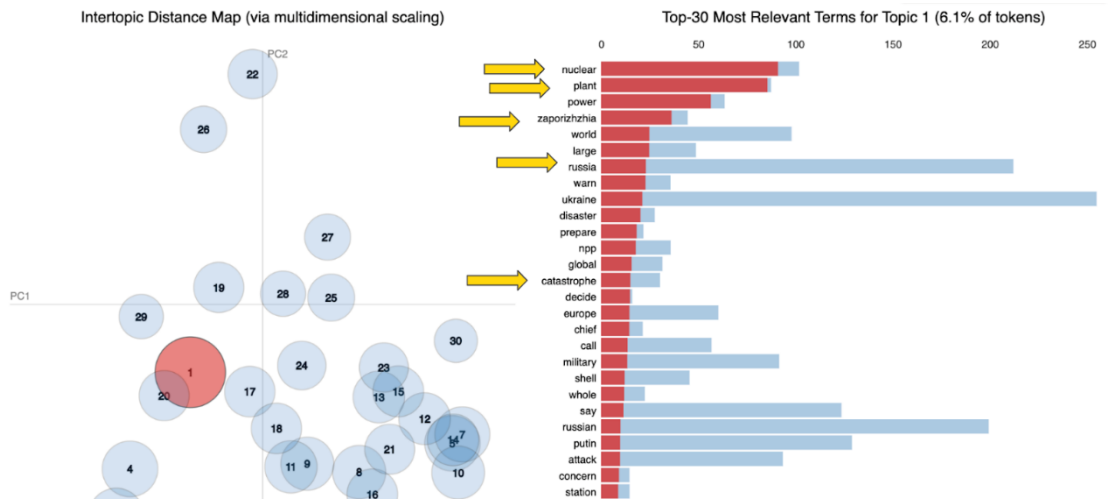

Figure 8. Twitter Negative Topic 1 - Zaporizhzhia Nuclear Plant

Factor 3 - Russia launches the missile attack on residential areas of Ukraine

This is identified as Topic 2 from Twitter's negative posts and Topic 11 from Reddit's negative posts. Keywords like kill, child, missile, Kharkiv, civilian, lives, etc. This topic implies that many innocent Ukrainians living in Kharkiv were killed, and the city broke down because Russia's missile strike hit schools and residential buildings. [3]

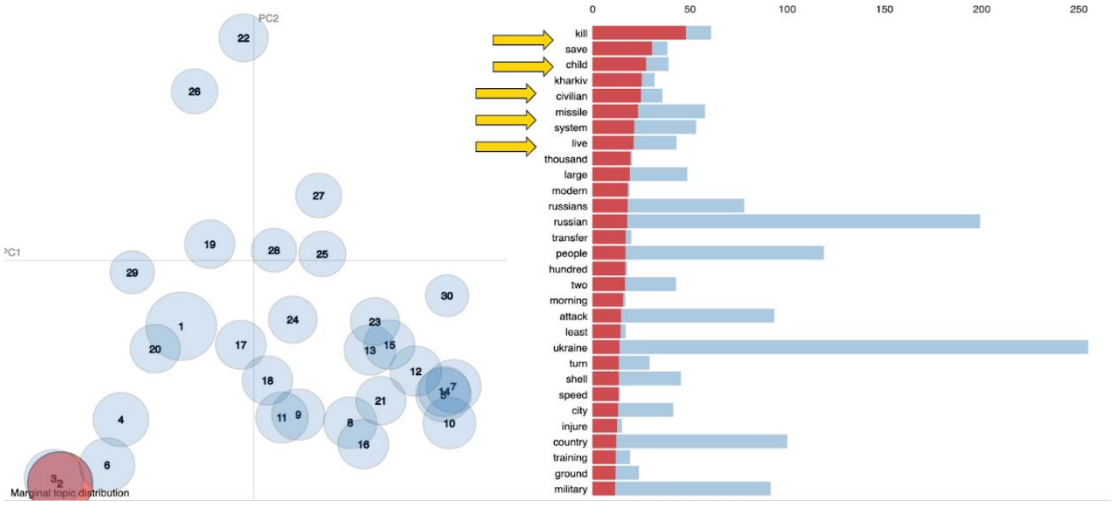

Figure 9. Twitter Negative Topic 2 - Civilian Massacre in Ukrainian Cities

Factor 4 - Putin's Regime

This is identified as Topic 19 from Twitter's negative posts. Vladimir Putin's regime had become a pariah state. The invasion of Ukraine by Russia forced the Europeans to change how they viewed threats. They concluded that the Russian threat posed a greater security risk to Europe than terrorism. [4]

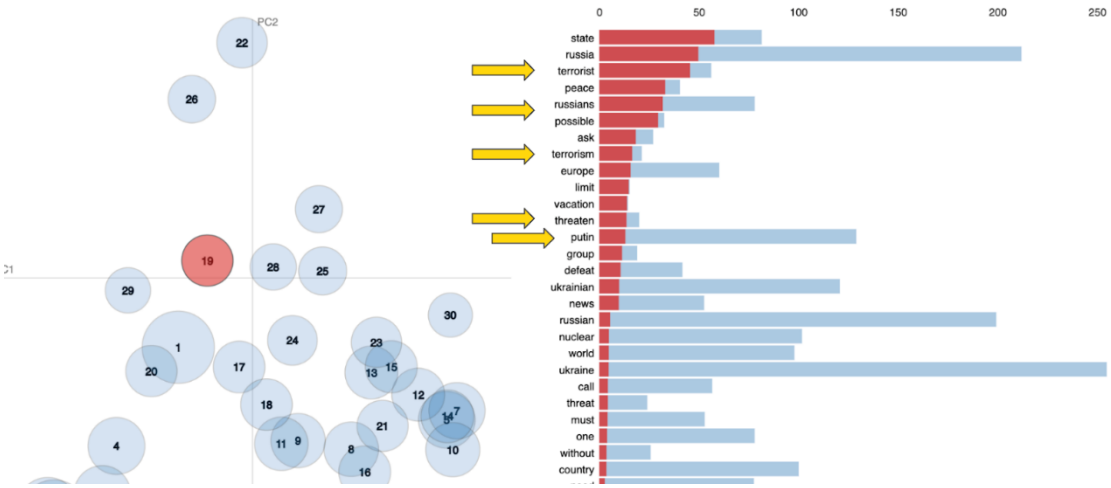

Figure 10. Twitter Negative Topic 2 – Terrorism Accuse against Putin

Factor 5 - Crimean Bridge Explosion

This is identified as Topic 4 from Twitter's negative posts. There are keywords like Crimea, explosion, Russian, attack, and bridge. The bridge is used to supply Russian troops in occupied Crimea during the invasion of Ukraine. Therefore, Russia has been accusing Ukraine of the explosion. This topic harms Ukraine's international presence and image. [5]

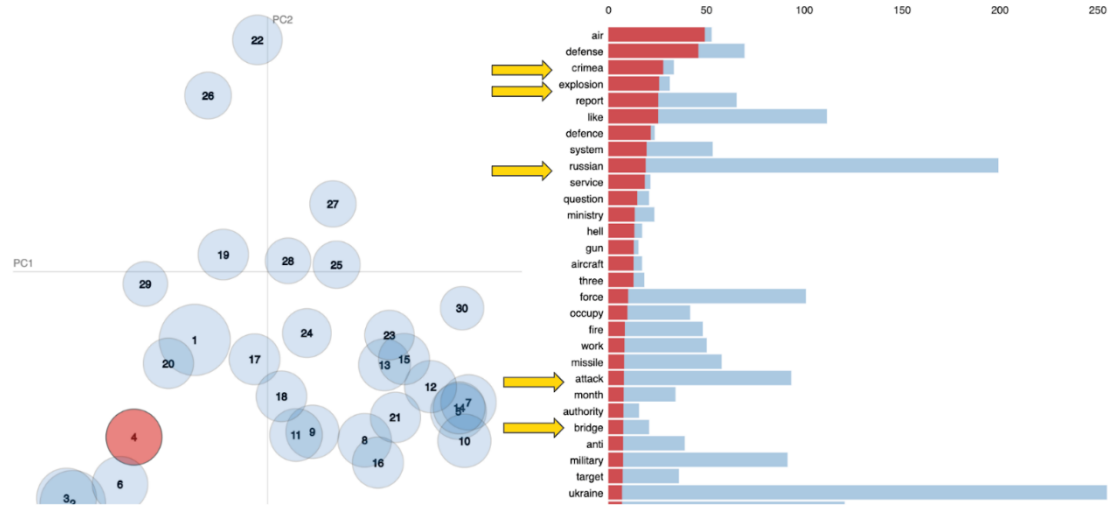

Figure 11. Twitter Negative Topic 4 – Crimean Bridge Explosion

4. Recommendations

As shown in the word cloud of Tweets with a negative sentiment, the majority of people on social media demonstrate a negative emotion about the war, especially for topics regarding killing innocent civilians, destroying public facilities, Nazi behaviors, etc.

Figure 12. Word Cloud of Tweets with a Negative Sentiment

On the other hand, the theme of the minority positive sentiments as shown in the word cloud of Reddit with a positive sentiment is about hope, love, blessing, support, etc.

Figure 13. Word Cloud of Reddit Posts with a Negative Sentiment

This finding suggests that people are born with conscience and empathy, and Russians are no different. The Russian soldiers likely are either blindfolded, misled, or brainwashed by the Russian government to fight for their country over so-called “national security issues due to the threat from the Western countries”. Therefore, we

want to propose the following recommendations to the Ukraine government and NGOs for defeating the invasion of Russia.

Recommendation 1 - Let Russians know the truth about the war

Ukrainian government or NGOs may want to create documentaries about the evil truth of the war, the abuse of humanity, and the desperate situation of Ukrainians. Then, they should infiltrate and propagate this documentary on Russian social media. Such that Russian citizens may have a chance to know that their government has been committing serious war crimes, and their sons, parents, or friends are sacrificing their lives for nothing but the Nazi ambitions of Vladimir Putin. It is never too hard to echo the hatred against Nazi crimes among Russians, as they stood by the correct side in World War 2. Again, the best way to collapse the morale of the Russian army is from the inside, and a proper documentary should be strong enough to awaken the conscience and the humanity of those Russian soldiers.

Recommendation 2 - Keep Hope and Provide Support

Although the past couple of months have been devastating when innocent Ukrainian were being slaughtered, we can't ditch the most powerful weapon we have which is hope. NGOs should urge more countries to join the fight for peace and justice and to support Ukraine as much as they can. It's vital to help Ukrainians to see the light of hope even in the darkest days. Furthermore, we shall let more people from around the world realize the negative impact of the war and direct them on a path to donate and assist the people in the war zones and to support the post-conflict reconstruction in Ukraine.

5. Conclusion and Call for Action

Upon performing sentiment classification of the social media posts/tweets about Russia's war in Ukraine, we notice that although Ukraine is still in a challenging situation, more and more people have realized who is guarding democracy, freedom, and the future of humanity while who is Nazism and the destroyer of peace. We believe that Ukraine is staying in a favorable public opinion environment. The Ukraine government/NGO should consider our recommendations and take advantage of the power of public opinion to awaken compassion, build consensus on justice, and unveil the evil truth of the war to every corner of the world.

6. Reference

[1] Nord Stream 1: How Russia is cutting gas supplies to Europe. BBC. Retrieved from: https://www.bbc.com/news/world-europe-60131520

[2] Hutchinson & Wholf, B.H. &T.W. Fighting around Ukraine's nuclear plants raises global concerns. ABCNEWS. Retrieved from: https://abcnews.go.com/International/fighting-ukraines-nuclear-plants-raises-global-concerns/story?id=83211363

[3] Deadly Russian Missile Strikes Hit Kharkiv As Death Toll Rises To 31 In Destroyed Apartment Block. REGIONS. Retrieved from: https://www.rferl.org/a/ukraine-chasiv-yar-russia-attack-civilians-killed-invasion/31937669.html

[4] Motyl, A.M. Ukraine War: Vladimir Putin has gambled everything and lost. Atlantic Council. Retrieved from: https://www.atlanticcouncil.org/blogs/ukrainealert/ukraine-war-vladimir-putin-has-gambled-everything-and-lost/

[5] Sackur, L.S. Huge explosion destroys part of bridge linking Russia and Crimea. NBC News. Retrieved from: https://www.nbcnews.com/news/world/russia-ukraine-war-explosion-kherson-bridge-crimea-putin-birthday-rcna51324