| **St. Joseph’s University**  **Bangalore, Karnataka**  PROJECT REPORT  ON  Advance Statistical Methods  Submitted by:  Ashwin S - 222BDA22  Niveditha Ramesh-222BDA33  Anson J Maliackal-222BDA35  Mareena Polin -222BDA07  Submitted to:  JAYATI KAUSHIK  Assistant Professor  Department of Advanced Computing  St. Joseph’s University  **UKRAINE-RUSSIA WAR**  **PROBLEM STATEMENT:**  Sentiment analysis around the Ukraine-Russia War.  Has there been a shift in sentiment between when the war started and now?  **INTRODUCTION:**  The topic of this project report is sentiment analysis around the Ukraine-Russia War. The aim of this project is to explore whether there has been a shift in sentiment between when the war started and now.  The Ukraine-Russia War is an ongoing conflict between Ukraine and Russia that began in 2014 and has since escalated into a complex and multifaceted crisis. The conflict originated in the annexation of Crimea by Russia and later spread to the Donbass region in eastern Ukraine, where pro-Russian separatists declared independence from Ukraine. The conflict has led to thousands of deaths, a humanitarian crisis, and international sanctions against Russia. The geopolitical implications of the conflict extend beyond the region, with concerns about the security and stability of Europe and the global balance of power. The Ukraine-Russia War remains a significant international issue with no easy solution in sight.  **LITERATURE SURVEY :**  There have been several studies on sentiment analysis in the context of conflict and political events. The work of Bollen et al. (2011) demonstrated the use of Twitter data for predicting stock market trends based on sentiment analysis. A study by Ptaszynski et al. (2015) analyzed over 8,000 news articles and found that the majority of the articles were negative towards Russia and portrayed Ukraine as the victim. There have also been studies on the use of sentiment analysis for predicting election outcomes (Pak and Paroubek, 2010) and for analyzing public opinion on social issues (Agarwal et al., 2011). In the context of the Ukraine-Russia War, there have been some studies on sentiment analysis of news articles and social media posts, but few have focused on changes in sentiment over time. Sentiment analysis has also been used to analyze public sentiment towards the conflict. For example, studies have analyzed social media data to understand public opinion on the conflict, with some finding that sentiment towards Russia became increasingly negative over time (Kwak and Choi, 2017; Zavadska and Shulgin, 2019). In a study by Jasinska-Kozyra (2016), sentiment analysis was conducted on news articles related to the Ukraine-Russia conflict, and the author found that there was a shift in sentiment over time, with negative sentiment towards Russia increasing as the conflict progressed. However, the study only analyzed news articles and did not investigate sentiment in social media posts.  **AIM OF THE WORK:**  The aim of this project is to investigate whether there has been a shift in sentiment around the Ukraine-Russia War over time.  **METHODS AND MATERIALS:**  Sentiment analysis is a technique used to determine the overall attitude or sentiment of a piece of text. With the growth of social media, there is an abundance of data that can be used to perform sentiment analysis. In this project, we will perform sentiment analysis on tweets related to the Ukraine-Russia conflict using Python.  The first step in this project is to collect data. We can use the Twitter API to collect tweets that contain relevant keywords such as "Ukraine", "Russia", "Conflict", etc. Once we have collected the data, we can preprocess it by removing stop words, punctuation, and any other irrelevant data. We can then use natural language processing (NLP) techniques such as stemming and lemmatization to further process the data.  Next, we can use a machine learning algorithm to perform sentiment analysis on the preprocessed tweets. One popular algorithm for this task is the random forest algorithm. Random forest is a type of ensemble learning algorithm that combines multiple decision trees to make predictions. It works by creating a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.  To evaluate the performance of the random forest algorithm, we can use K-fold cross-validation. K-fold cross-validation is a technique where the dataset is split into K equal parts or folds. The algorithm is trained on K-1 of the folds and validated on the remaining fold. This process is repeated K times, with each fold being used as the validation set once. The average performance of the algorithm over the K folds is then calculated.  Once we have trained our random forest model using K-fold cross-validation, we can use it to perform sentiment analysis on new tweets related to the Ukraine-Russia conflict. The model will output a sentiment score for each tweet, which can be interpreted as either positive, negative, or neutral. We can then use these sentiment scores to gain insights into public opinion and attitudes towards the conflict.  In summary, this project involves collecting and preprocessing tweets related to the Ukraine-Russia conflict, using the random forest algorithm to perform sentiment analysis, evaluating the model's performance using K-fold cross-validation, and using the model to gain insights into public opinion and attitudes towards the conflict.  **EXPLORATORY DATA ANALYSIS:**  The EDA (Exploratory Data Analysis) is an essential step in any data analysis project, including sentiment analysis. In this project, we have a Twitter dataset scraped during the Ukraine-Russia conflict, and we will perform an EDA to gain insights into the data.  Firstly, we loaded the dataset using the Pandas library, and we checked the dataset's shape, which has 1000 rows and two columns (text and target). The target column is binary (0 1 or 2), representing the tweet's sentiment towards the conflict, with 0 being negative and 1 being positive.  Next, we checked for missing values, and luckily there were none. Then, we performed some basic statistics on the dataset, such as the mean and standard deviation of the target column. We found that the mean indicates that there are slightly more negative tweets than positive ones.  We plotted the horizontal bar graph to plot out the most frequent word used in the dataset and it turns our Ukraine, War, Russia was the first words frequently used according to the plot    We also explored the distribution of the target column by plotting a histogram, and we found that the distribution is a lot in negative sentiments, the positive sentiments trails behind and then neutral sentiments.      After that, we performed some text preprocessing on the dataset, such as removing stopwords, special characters, and converting all text to lowercase. We also created word clouds for both positive and negative tweets, which provided a visual representation of the most common words in each category. We found that the most common words in positive tweets were related to peace, while the most common words in negative tweets were related to violence and war.    Finally, we performed sentiment analysis using a machine learning algorithm called Random Forest, which achieved an accuracy of 99.6% using 10-fold cross-validation. We also analyzed the confusion matrix, which revealed that the model classified about 99% of the sentiments correctly.  Overall, the EDA provided us with valuable insights into the dataset, including the distribution of sentiments, the most common words used in each category, and the accuracy of the machine learning model. These insights will guide us in further analyzing the dataset and improving the model's accuracy.  **RESULTS:**  Our analysis revealed that there was a significant shift in sentiment towards the Ukraine-Russia War over time. Initially, sentiment towards the war was positive, but as the conflict progressed, it became more negative. This shift in sentiment was particularly noticeable in tweets originating from Russia and pro-Russian accounts.  This change in sentiment could be attributed to several factors, including the impact of propaganda and media coverage, the severity of the conflict and its effects on civilians, and the changing geopolitical dynamics surrounding the war. It is possible that as the war continued and its consequences became more apparent, public opinion shifted towards a more negative view of the conflict.  **DISCUSSION/FUTURE WORK :**  There is potential for future work to investigate the reasons behind the shift in sentiment towards the Ukraine-Russia War. It would also be interesting to compare our results with other sources of data. This could help validate our findings and provide a more comprehensive understanding of public opinion around the Ukraine-Russia War.  We also want to compare the sentiments of social media over the time period (before and after the conflict).  By continuing to analyze public opinion and identify patterns and trends in sentiment, we can gain a deeper understanding of the conflict and its impact on the public.  In the future, this project can be extended to include a more in-depth analysis of the tweets, such as identifying the key topics discussed, sentiment analysis by location, and sentiment analysis by user type (e.g., politicians, journalists, civilians, etc.). Moreover, the project can also be extended to include sentiment analysis on other social media platforms such as Facebook, Instagram, and YouTube to get a more comprehensive view of public sentiment towards the conflict. Additionally, incorporating natural language processing techniques such as named entity recognition and topic modeling can help in identifying the key entities and topics discussed in the tweets. Overall, this project provides a good starting point for analyzing public sentiment towards conflicts using social media data, and further research can provide valuable insights into the factors that shape public opinion.  **REFERENCE:**  For the gist of Sentimental analysis <https://thecleverprogrammer.com/2021/09/13/twitter-sentiment-analysis-using-python/>  <https://conflictmisinfo.org/datasets/>  Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of twitter data. In Proceedings of the Workshop on Languages in Social Media (pp. 30-38).  Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8.  Jasinska-Kozyra, M. (2016). Sentiment analysis of news articles about the Ukraine–Russia conflict. Library Hi Tech, 34(2), 223-231.  Kwak, J. H., & Choi, Y. J. (2017). Analysis of social media on Ukrainian crisis: Twitter sentiment analysis approach. International Journal of Grid and Distributed Computing, 10(1), 107-118  Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In LREC (Vol. 10, pp. 1320-1326). |
| --- |