A sentiment analysis of Reddit comments on the Ukraine-Russia conflict using exploratory data analysis and NLP

Abstract: Through exploratory data analysis of datasets, summarizing their main characteristics, and visualizing the collected data, this paper would give useful information about how the public feels about the Russian-Ukrainian war and about cultural differences and similarities.

Key terms: Russian-Ukrainian war, NLP, data analysis, VADER, histogram

I Introduction

During this research project, I intended to conduct a sentiment analysis on data scraped from social media site Reddit pertaining to the Russian-Ukrainian war. This sentiment analysis was to be done in order to gain an understanding of the collective opinion of Reddit users towards the war. Sentiment analysis is a type of natural language processing that extracts and quantifies the emotional content of text data using computational methods and assigns a score to the sequence of characters as either positive, neutral, or negative. The Russian-Ukrainian war subject was chosen because of its wide exposure on various social media platforms and communication channels and its relevance to current events. This information can be utilized to further our understanding of the general sentiment and find interesting insights on how people and nations are reacting to the war. I decided to focus specifically on one social network, Reddit, as it facilitates the posting of opinions via simple comments around the theme. Reddit allows an easy and anonymous way to post opinions, which can yield more accurate and honest opinions from the audience.

II Methods and Data

The first step is gathering the required sample data from the internet about the theme and then cleaning and structuring the obtained raw data for further exploratory data analysis. I identified the potential subreddits, which contain the public's opinions and could be a representative sample of cultural sentiment. In the beginning of my research, I utilized the NLP pipeline (VADER to assign the score and the NLTK package) and practiced with one blog post's comments. I then extrapolated the results and applied the pipeline to scrape data and comments from the entire

subreddit in order to achieve a more comprehensive representation. The more data we have, the larger the sample is, the more accurate the results. With the Python Reddit API wrapper (PRAW), I scraped 1048 comments from the "UACommunity" subreddit, which contains comments from three cultural groups and contains sufficient data in three languages. To compare three cultural groups, I had to divide data into three separate pandas data frames using the "langdetect" library, dividing each comment based on the language tags ('en', 'uk', and 'ru'). I also encountered difficulties with the original VADER library and its alternative TextBlob during my exploration, as they did not support the "Russian" and "Ukrainian" languages and only supported "English." Therefore, after researching, I identified the "VADER Sentiment Analysis Multilanguage". Using this particular library, I then processed the unstructured data to obtain sentiment scores for the comments in all three languages and classified them as positive, negative, or neutral.

index	neg	neu	pos	compound	headline	language
0	0.000	1.000	0.000	0.0000	туалет для Рогозина	ru
1	0.000	0.897	0.103	0.3182	Документы, пожалуйста. Где и сколько раздали п...	ru
4	0.108	0.808	0.084	-0.3612	На России опять ебнулись и в рекламе спокойно ...	ru
5	0.307	0.693	0.000	-0.8519	Запорожье переживает последствия новогодней вр...	ru
6	0.000	1.000	0.000	0.0000	Кто знает куда обратиться, что бы началась про...	ru
...	...	...	...	...	...	...
974	0.193	0.807	0.000	-0.8834	В Ростовской области ищут шестерых сбежавших и...	ru
975	0.000	1.000	0.000	0.0000	"Партизанский Телеграм", в котором есть инстру...	ru
976	0.038	0.853	0.109	0.5980	В Италии прошел перфоманс в поддержку сбора де...	ru
977	0.000	1.000	0.000	0.0000	Вышло новое интервью на канале Золкина	ru
978	0.224	0.561	0.215	-0.0258	Адмирал Кузнецов загорелся во время ремонта... с...	ru

index	neg	neu	pos	compound	headline	language
51	0.000	1.000	0.000	0.0000	Надо определяться	uk
77	0.000	1.000	0.000	0.0000	ПРЕДЗАКАЗ	uk
134	0.317	0.509	0.174	-0.7003	Луганчанин Иванов (журналист) и азовец ТАВР: п...	uk
312	0.187	0.620	0.193	0.0191	Давно я так не ржал..	uk
330	0.000	0.000	1.000	0.4019	Да	uk
364	0.191	0.564	0.244	0.4003	Вічна пам'ять загиблим героям! ❤️Герої, які по...	uk
406	1.000	0.000	0.000	-0.6124	Ублюдки.	uk
507	0.000	0.303	0.697	0.3182	ПОЖЕЛТЕВШАЯ ПРАВДА	uk
566	0.000	0.481	0.519	0.4995	TARAS KEEN - Zомбі	uk
580	0.188	0.673	0.139	-0.2500	«Дождь»: чи достатньо російському медіа бути п...	uk
837	0.146	0.704	0.150	0.0258	Законопроєкт про посилення відповідальності дл...	uk

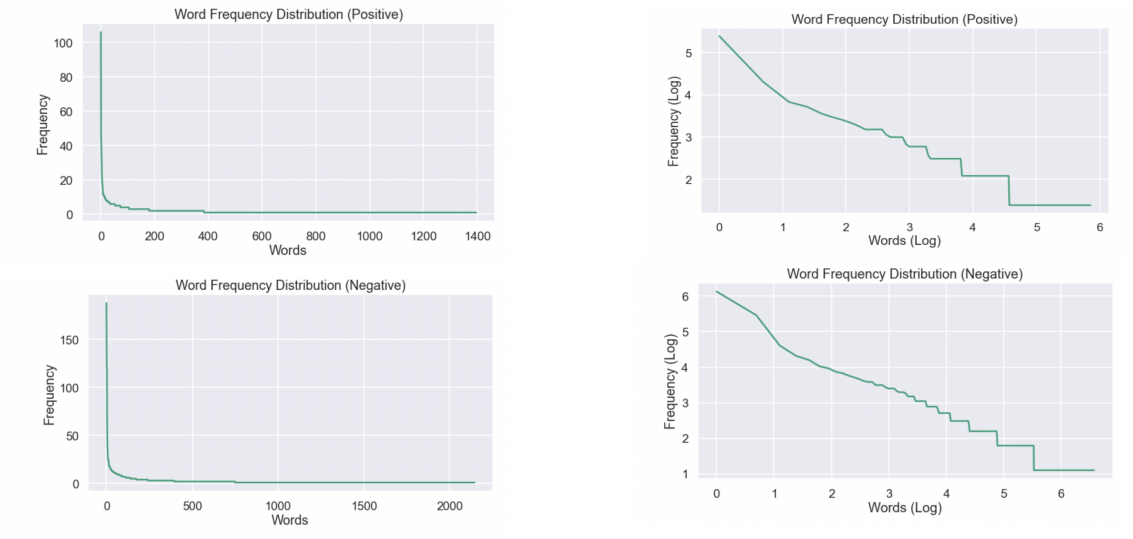
index	neg	neu	pos	compound	headline	language
2	0.000	0.716	0.284	0.9018	Russian Volunteer Corps 🪚☀️: *Friends! Today, a...	en
3	0.000	0.488	0.512	0.6369	The Greatest Beer Run Ever	en
10	0.042	0.842	0.115	0.4588	Russian Volunteer Corps 🪚☀️: *As we wrote earli...	en
18	0.000	0.791	0.209	0.6696	The story of a mural with Putin in Belgrade, S...	en
33	0.112	0.745	0.144	-0.0258	Russian Volunteer Corps posts a unique intervi...	en
...	...	...	...	...	...	...
931	0.175	0.825	0.000	-0.7845	Feminist AntiWar Resistance 🤷‍♀️: *A member of...	en
935	0.570	0.430	0.000	-0.6486	Russian Partisan Groups Attacking Critical Inf...	en
936	0.000	1.000	0.000	0.0000	The Green Gendarmerie is recruiting volunteers...	en
938	0.217	0.783	0.000	-0.3612	Belarus legalizes pirated movies, music and so...	en
954	0.000	0.873	0.127	0.4404	Russian Volunteer Corps 🪚☀️: Our article: *The ...	en

Three tables represent the data frames for separate cultural groups; "headline" is the comment in the corresponding language, and "language" label content is the tag: 'en,' 'ru', 'uk' respectively.

Tokenization

I also proceeded to tokenize and identify the most frequent words by utilizing the NLTK package's tokenizer class and word tokenizer's RegexpTokenizer to extract tokens and character sequences. In my dataset, using those tools, I identified the "stop words," which are typically

common words present in a language that add little to the meaning of the text. Consequently, I removed them from the NLP tasks and continued tokenization and analysis of the rest, generating charts for the most frequent words. In my investigation, the initial results (diagrams on the left) are nearly identical, making it difficult to determine which words are most frequent.

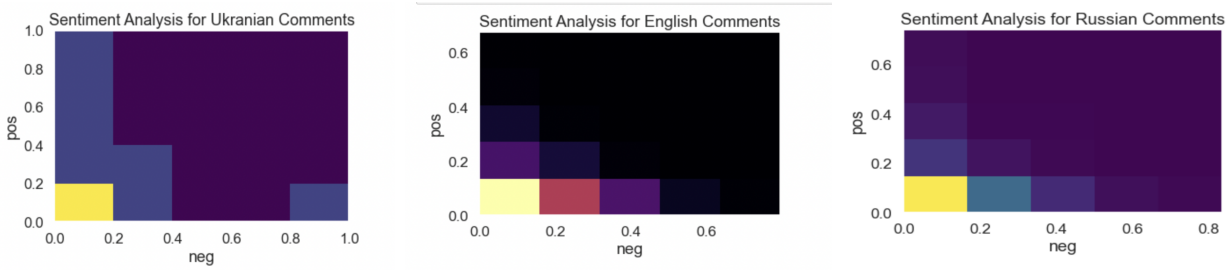


Therefore, I used logarithmic time instead, because it allows for more meaningful data visualization and analysis. Because the number of words in the dataset was large, the frequency distribution was difficult to interpret since frequent words appear more often than less frequent ones, as illustrated in diagrams on the left, which show a large, steep decline for both positive and negative sets. In contrast, the logarithmic scale allows for some pattern recognition, and the most frequently occurring words tend to cluster on the left side of the graph and then uniformly decline. However, it is important to note that the limitations of my data are due to the fact that VADER assigned a score to the comments as a whole, and then I tokenized the data, which produced a lot of false positives. In future work, duplicate values, such as "Ukraine," "Russia," "war," and "attack," and others which appear in both sets (positive and negative), would have to be filtered, producing more accurate results and allowing us to learn more deeply about differences of words, and clearly define the most frequent negative and positive words. At the same time, the significance of context and variations in word meanings could pose a problem, as each word can be interpreted differently based on the context in which it is used and culture. Thus, the limitations of my data can be addressed with improved techniques such as filtering and accounting for context and culture in the interpretation of words, potentially defining the most

frequent positive and negative words separately for three cultures, and then performing analysis and identifying trends and similarities.

III Histograms

By evaluating the data, I discovered that a considerable proportion of the comments had a neutral sentiment score, making it impossible to acquire a clear understanding of the variations and similarities in sentiment amongst the three ethnic groups. To solve the problem of neutral score majority, I constructed a histogram in which each comment was represented as a dot on a histogram based on whether it was positive or negative. Matplotlib and the hist2d function were used to create histograms, with negative sentiment ratings on the x-axis and positive sentiment scores on the y-axis. However, prior to beginning of analysis, it is necessary to understand the significance of color and its intensity. When a high number of data points cluster together, the color gets more strong and concentrated; otherwise, it becomes lighter and less vivid.

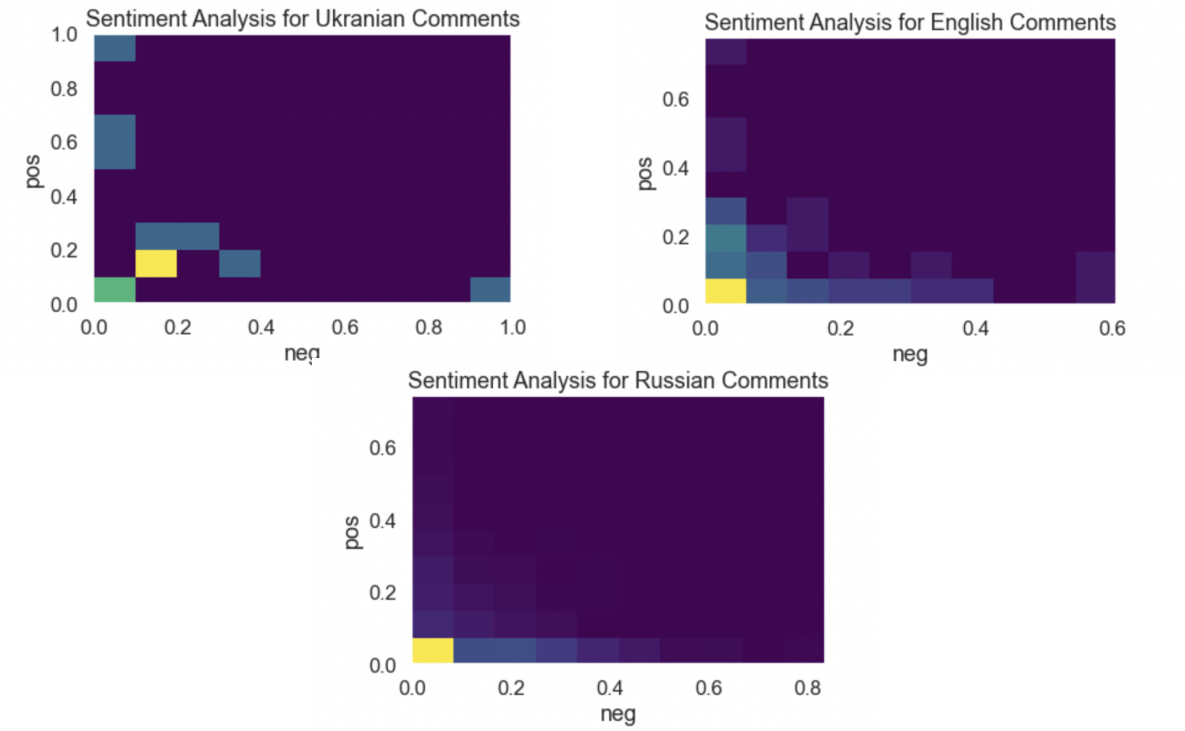


Histograms figures for sentiment analysis of three cultural groups: Ukraine, English and Russian

Function to create a histogram: ***plt.hist2d(x, y, bins=5, range=None, density=False, weights=None, cmin=None, cmax=None, data=None, cmap='viridis')***

To examine my data, I also changed the value of the function's parameter "bins" from 5 to 10. When the number of bins in the histogram increases, the histogram has a higher resolution and displays more detail. When I used the value 5 for the bins, it was difficult to tell the difference between the English and Russian comment representations, they were almost identical. Therefore

by increasing it to 10, histograms now reflect a narrower range of values, which allows for more detailed observation of fluctuations in the data distribution.



IV Analysis and Interpretation

I would first check the data distribution focusing on the density and shape of the data points, before analyzing the histograms. In Ukrainian comments, the distribution of emotion ratings was skewed toward the positive end of the scale, with a higher proportion of comments scoring positively. Ukrainian comments are divided into three high-density (intensely colored) regions: some are concentrated in the lower-left, some in the upper-left, and some in the lower-right. Ukraine is the only culture with a thick lower-right zone, signifying a positive collective sentiment. The English comments data points are uniformly distributed across the histogram, indicating that the data is balanced between positive and negative sentiment and has less intent. The majority of the English remarks are in the lower left, but they are evenly spread, with a few less noticeable locations in the lower right. The majority of the data points for the Russian comments are in the lower left corner of the histogram.

Overall, histogram visualizations reveal a significant cultural difference in online debates about the Russia-Ukraine war. Because the lower-right corner of the histogram has a high density, it demonstrates that Ukrainian comments have a relatively high proportion of positive emotion, which is unique compared to other cultures. Thus we can assume that Ukrainian views are more hopeful about the war compared to other cultures. The equally dispersed data points across the histogram indicate that the sentiment in English comments is largely balanced. As a result, the majority of English remarks on the war are neutral, with nearly equal proportions of favorable and negative opinions. Finally, since the majority of data points in the lower-left quadrant of the histogram, Russian comments have a higher proportion of negative sentiment. Russian remarks on the war are more pessimistic than English and Ukrainian comments. It is also worth noting that, while Ukrainian comments contain a pretty high proportion of positive sentiment, they also contain a relatively high proportion of negative sentiment, as evidenced by the density of data points in the lower-left corner. Ukrainian reactions to the war are varied, with both positive and negative feelings expressed.

V Conclusion

To conclude, this research results highlight the emotional similarities and differences between the three cultures' sentiments. Taking the high overview, the Ukrainian comments are more positive and nuanced, while English comments are neutral and balanced, and Russian comments are more negative. Despite the fact that Ukrainians have largely been on the receiving end of the conflict, their responses to the war show a complex mixture of emotions. Ukrainians feel both positive emotions such as patriotism and pride, as well as negative ones such as frustration and sadness. These nuanced responses among Ukrainians showcase that the conflict is much more complicated than a simple binary between “us” and “them.” This can be contrasted with the responses of Russians, who express almost exclusively negative emotions like, aggression and mistrust. Therefore, while the English comments provide a more neutral and balanced viewpoint, as they are not directly involved in the conflict, the contrast between the Ukrainian and Russian responses clearly indicates that emotions are playing a big part in the conflict and war. The difference in the responses to the conflict among Ukrainian, English, and Russian citizens illustrate the diversity of human emotion. In the end, this analysis has demonstrated that the cultural context has a significant impact on public opinion regarding the conflict in Ukraine.

Reference List:

Bajaj, Aryan. “VADER Sentiment Analysis | NLP Sentiment Analysis Using VADER.” *Analytics Vidhya*, 17 June 2021, www.analyticsvidhya.com/blog/2021/06/vader-for-sentiment-analysis/.

López Ramírez, Isaac and Méndez Vargas, Jorge. "A Sentiment Analysis of the Ukraine-Russia Conflict Tweets Using Recurrent Neural Networks." (2022).

Martínez, Rodrigo. “VADER Sentiment Analysis Multilanguage.” *GitHub*, 28 May 2022, github.com/brunneis/vader-multi. Accessed 21 Jan. 2023.

Sus, Lukasz. “Introduction to Sentiment Analysis in NLP.” *Www.netguru.com*, 22 July 2022, www.netguru.com/blog/sentiment-analysis-nlp. Accessed 21 Jan. 2023.