Title: Sentiments Surrounding the Russia-Ukraine Conflict Globally

Authors: Roshan Bellary, Tevin Wang, Spencer He, Winston Li

Introduction/Hypothesis:

The Russia-Ukraine conflict first began in 2014 with the Russian invasion of Crimea and Donbas where the former was annexed after a war broke out in 2022. Very recently however, the Russian-Ukraine war has ignited again with the Russian invasion of Donbas and continued fighting to the heart of Ukraine, Kyiv. These new conflicts have ignited a territorial fight between Ukraine and Russia which has led to 5.1 million Ukrainians and a quarter of the Ukrainian population being displaced.

In this modern era, there's another platform that facilitates an extremely important aspect of war. The Russia-Ukraine War isn't just being fought in Ukraine, but also online. Both the Russian and Ukrainian governments have utilized social media to try to shape public opinion in their favor, which can have a great effect on morale, public support, and possibly the course of the conflict. Public opinion matters to political leaders, military officials, and humanitarian aid workers. It can shape how the war is fought, how motivated soldiers are, where aid needs to be distributed most, and who is more likely to win.

Statistical Question:

This study will deal with the question, how does public perception of the Russia-Ukrainian conflict vary throughout the world, and what emotions or sentiments do citizens in these countries have with regards to the conflict?

We hypothesize that western European countries such as Germany, France and the UK along with North American countries such as the US and Canada will have significantly higher proportions of negative sentiments regarding the conflict, while countries that are more politically aligned with Russia such as Belarus and Armenia will have lower proportions of these sentiments. We also expect that countries that are closer to the conflict will have higher proportions of extreme emotions, such as fear, as the war is being fought near them. Additionally, we hypothesize that the main sentiments that will be exhibited by the former group of countries would lean toward anger and sadness while for the latter group it would lean more towards joy or neutrality.

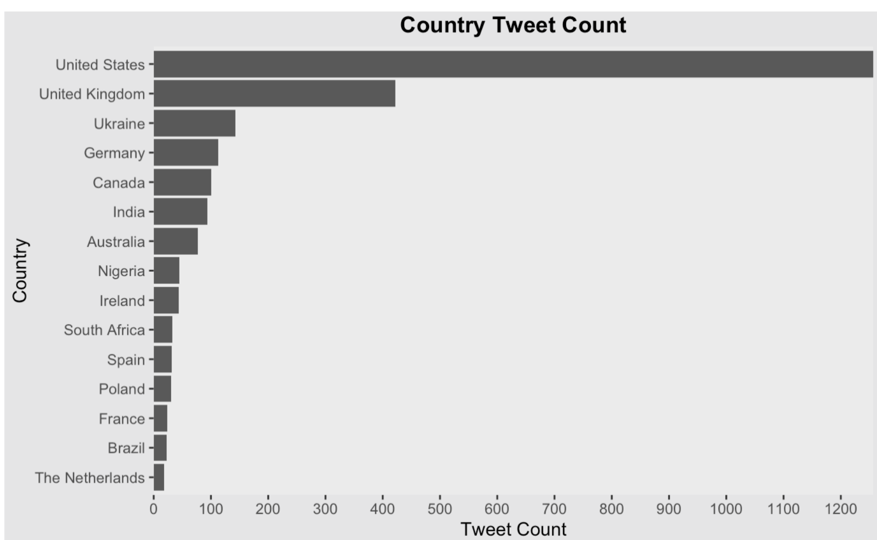
Studying the Data:

Our first problem was figuring out what type our data should be and where we should collect it. In order to get a baseline for the sentiments displayed in every country, we decided to use a social media site and collect data from the site pertaining to the conflict. Going off of this, we found a Kaggle dataset on tweets with certain keywords related to the war from January 1st to March 6th [1]. We decided to use the "ukraine war" keyword in order to narrow our analysis to specifically tweets with war and ukraine-related information.

Cleaning the Data:

In the dataset with the "ukraine war" keyword, there are over 200 thousand tweets over the 65 days of data collecting. However, we ran into an important problem. Tweets without location metadata (which occurs when someone turns off sharing their location with Twitter) would be unable to be analyzed as our analysis is based upon the location of the tweets. So we decided to prune our dataset of any tweets without location data. Our resulting dataset had approximately 2900 tweets all with the location from which they were sent [2]. The metadata included with the tweet had the following main variables-date, content of the tweet, user id along with corresponding information, language, latitude, longitude and finally region from which the tweets were sent(a city code along with a country code). Using this metadata along with the content of the tweet, we could now go forward with our analysis.

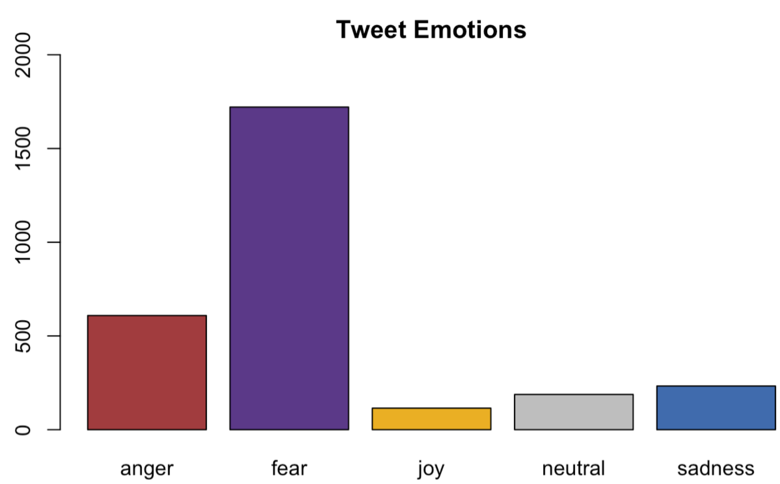
Wordcloud of words contained in dataset



# of tweets per country in dataset

Engineering the Features:

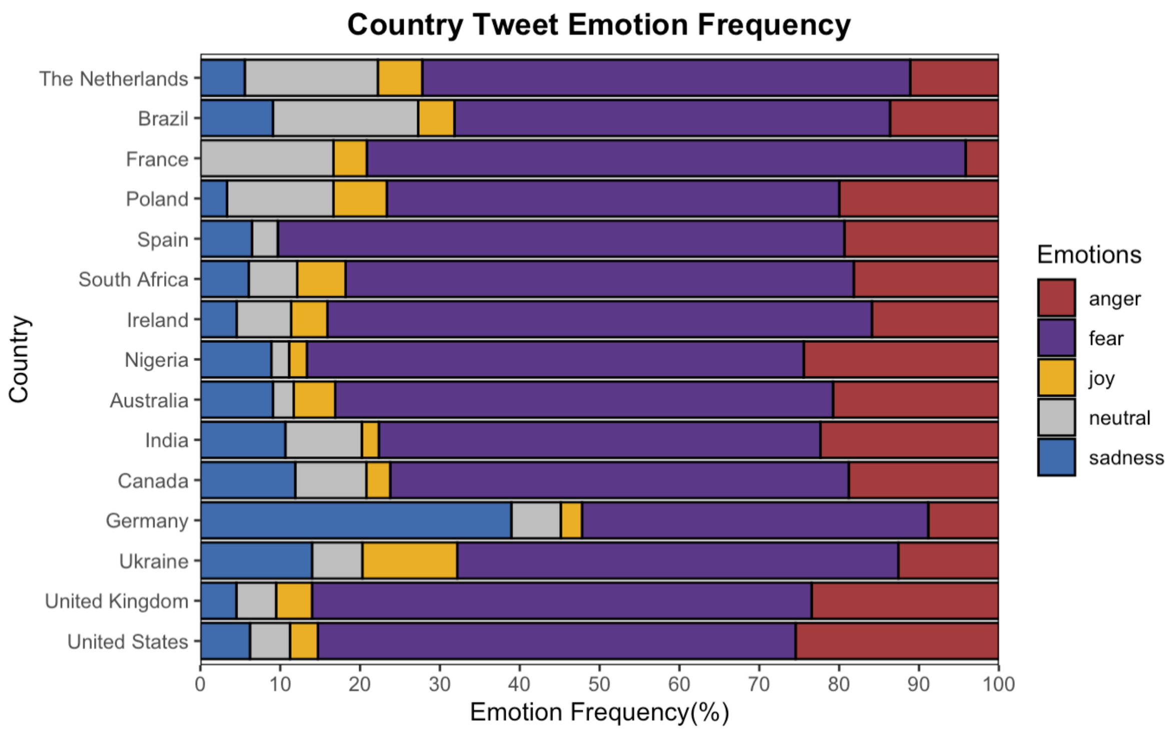
To determine the sentiment of a tweet, we decided to use sentiment analysis. Sentiment analysis works by breaking up a piece of written content into pieces of topic chunks and assigning a sentiment score to each topic chunk to detect the sentiment of the entire message. The particular model that we used turned words in a message into vectors then used inbuilt-classifiers with the sci-kit learn library in python to classify the message into 5 different emotion categories-joy, sadness, anger, fear and neutral [3]. We then applied the trained model to our collection of tweets with metadata and extracted a sentiment out of every tweet creating a new dataset with corresponding sentiments per each tweet.



# of Tweets per Emotion

Combining our location data with corresponding tweet emotions, we were also able to create a graph of the country's tweet emotion frequencies which showed the distribution of tweet sentiments in the 15 countries which had the most tweets in the dataset.

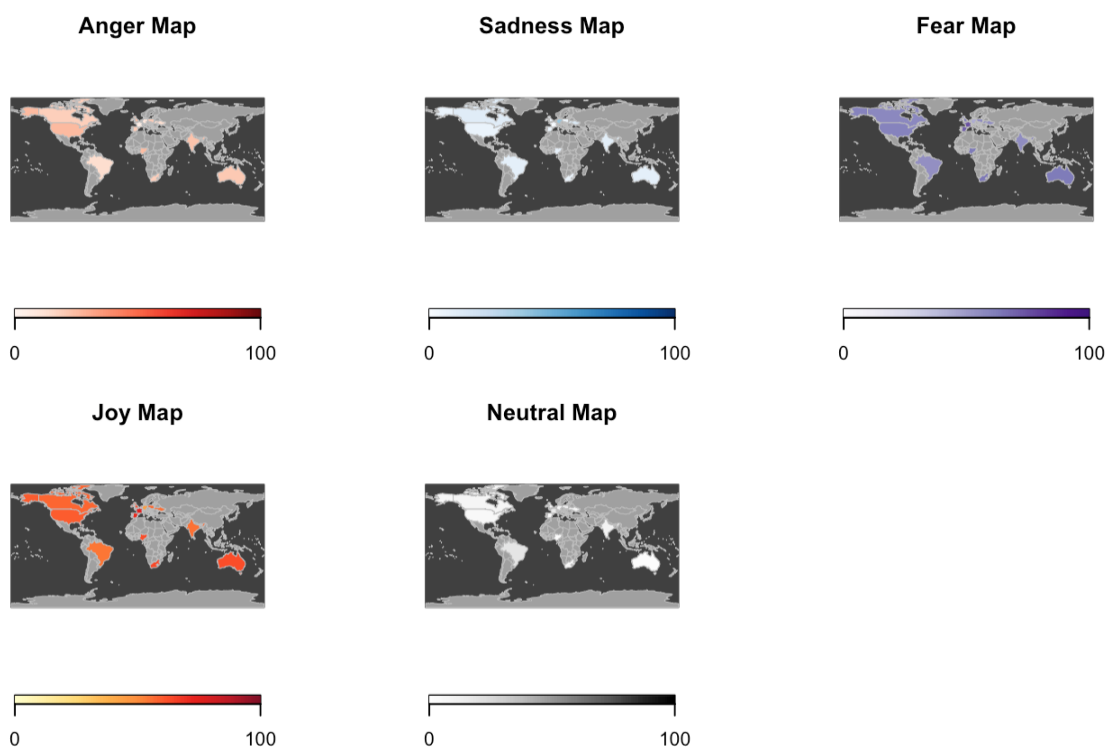
Frequency distribution of tweet emotions per country



Emotion distribution for top 15 countries(%)

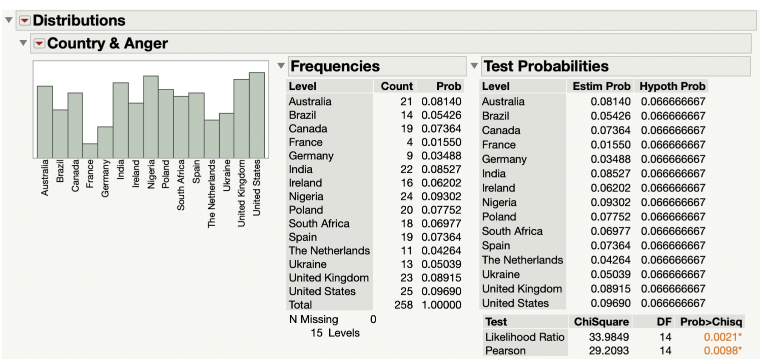
	Country	Anger	Fear	Joy	Neutral	Sadness
1	The Netherlands	11.1111111111111	61.1111111111111	5.55555555555556	16.6666666666667	5.55555555555556
2	Brazil	13.6363636363636	54.5454545454545	4.54545454545455	18.1818181818182	9.09090909090909
3	France	4.16666666666667	75	4.16666666666667	16.6666666666667	0
4	Poland	20	56.6666666666667	6.66666666666667	13.3333333333333	3.33333333333333
5	Spain	19.3548387096774	70.9677419354839	0	3.2258064516129	6.45161290322581
6	South Africa	18.1818181818182	63.6363636363636	6.06060606060606	6.06060606060606	6.06060606060606
7	Ireland	15.9090909090909	68.1818181818182	4.54545454545455	6.81818181818182	4.54545454545455
8	Nigeria	24.4444444444444	62.2222222222222	2.22222222222222	2.22222222222222	8.88888888888889
9	Australia	20.7792207792208	62.3376623376623	5.19480519480519	2.5974025974026	9.09090909090909
10	India	22.3404255319149	55.3191489361702	2.12765957446809	9.57446808510638	10.6382978723404
11	Canada	18.8118811881188	57.4257425742574	2.97029702970297	8.91089108910891	11.8811881188119
12	Germany	8.84955752212389	43.3628318584071	2.65486725663717	6.19469026548673	38.9380530973451
13	Ukraine	12.5874125874126	55.2447552447552	11.8881118881119	6.29370629370629	13.986013986014
14	United Kingdom	23.4597156398104	62.5592417061611	4.50236966824645	4.97630331753555	4.50236966824645
15	United States	25.4574383452665	59.8249801113763	3.50039777247414	5.01193317422434	6.20525059665871
16	Sum	259.08998525304	908.40574106791	66.6011346470722	126.734696223679	139.168442808299

Global Map of Emotion Frequencies



Data Analysis:

To test our hypothesis, we performed five Chi-square hypothesis tests — each corresponding to each emotion — to detect if there are any countries that have a statistically significant higher percentage of tweets of a certain emotion or not. Our null hypothesis for each emotion — anger, fear, joy, neutral, and sadness — was that the percentage of tweets with that emotion of each country would be equal. Our alternate hypothesis was that the percentage of tweets with that emotion of each country are different, which would indicate that there were certain countries that had a statistically different proportion of emotions compared to the others. We will be using a significance value of 0.05 for this test.



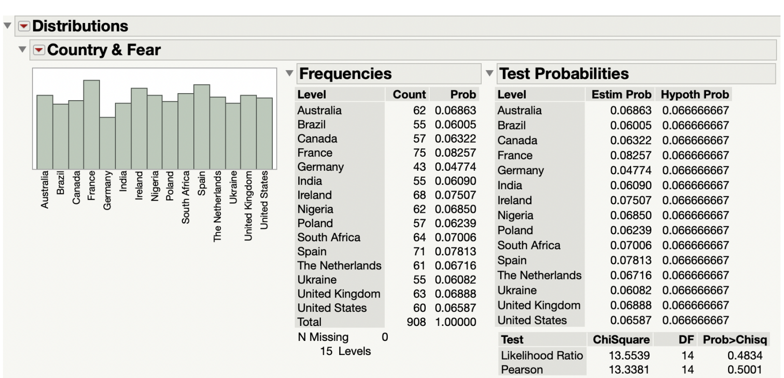
Level	Count	Prob
Australia	21	0.08140
Brazil	14	0.05426
Canada	19	0.07364
France	4	0.01550
Germany	9	0.03488
India	22	0.08527
Ireland	16	0.06202
Nigeria	24	0.09302
Poland	20	0.07752
South Africa	18	0.06977
Spain	19	0.07364
The Netherlands	11	0.04264
Ukraine	13	0.05039
United Kingdom	23	0.08915
United States	25	0.09690
Total	258	1.00000

N Missing 0
15 Levels

Level	Estim Prob	Hypoth Prob
Australia	0.08140	0.066666667
Brazil	0.05426	0.066666667
Canada	0.07364	0.066666667
France	0.01550	0.066666667
Germany	0.03488	0.066666667
India	0.08527	0.066666667
Ireland	0.06202	0.066666667
Nigeria	0.09302	0.066666667
Poland	0.07752	0.066666667
South Africa	0.06977	0.066666667
Spain	0.07364	0.066666667
The Netherlands	0.04264	0.066666667
Ukraine	0.05039	0.066666667
United Kingdom	0.08915	0.066666667
United States	0.09690	0.066666667

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	33.9849	14	0.0021*
Pearson	29.2093	14	0.0098*

Starting with anger, the graph of distributions shows France with a lower proportion of angry tweets compared to a country like the United States. The Chi-square test confirms this, as we receive a p value of 0.0021, less than 0.05, meaning we can reject our null hypothesis. There does appear to be a statistically significant difference among countries in terms of the proportion of tweets that correspond with anger.



Level	Count	Prob
Australia	62	0.06863
Brazil	55	0.06005
Canada	57	0.06322
France	75	0.08257
Germany	43	0.04774
India	55	0.06090
Ireland	68	0.07507
Nigeria	62	0.06850
Poland	57	0.06239
South Africa	64	0.07006
Spain	71	0.07813
The Netherlands	61	0.06716
Ukraine	55	0.06082
United Kingdom	63	0.06888
United States	60	0.06587
Total	908	1.00000

N Missing 0
15 Levels

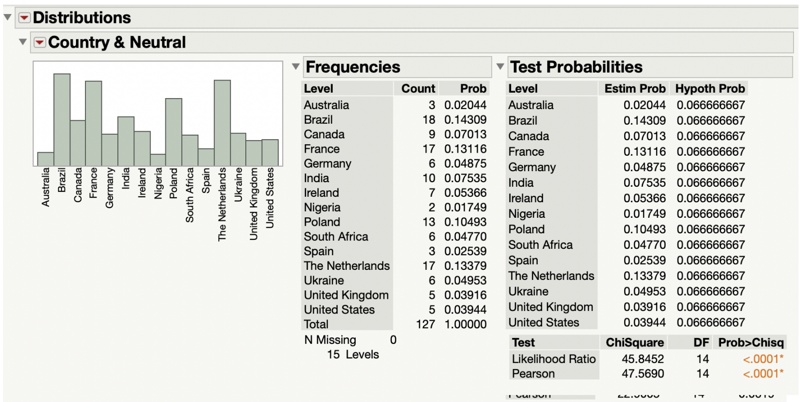
Level	Estim Prob	Hypoth Prob
Australia	0.06863	0.066666667
Brazil	0.06005	0.066666667
Canada	0.06322	0.066666667
France	0.08257	0.066666667
Germany	0.04774	0.066666667
India	0.06090	0.066666667
Ireland	0.07507	0.066666667
Nigeria	0.06850	0.066666667
Poland	0.06239	0.066666667
South Africa	0.07006	0.066666667
Spain	0.07813	0.066666667
The Netherlands	0.06716	0.066666667
Ukraine	0.06082	0.066666667
United Kingdom	0.06888	0.066666667
United States	0.06587	0.066666667

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	13.5539	14	0.4834
Pearson	13.3381	14	0.5001

For fear, the graph of distributions shows most of the countries have a relatively equal proportion of tweets that contain fear. The

Chi-square test confirms this, as we receive a p value of 0.4834, greater than 0.05, meaning we fail to reject our null hypothesis. There does not appear to be a statistically significant difference among countries in terms of the proportion of tweets that correspond with fear.

For joy, the graph of distributions shows some isolated differences in joy among the proportion



Frequencies

Level	Count	Prob
Australia	3	0.02044
Brazil	18	0.14309
Canada	9	0.07013
France	17	0.13116
Germany	6	0.04875
India	10	0.07535
Ireland	7	0.05366
Nigeria	2	0.01749
Poland	13	0.10493
South Africa	6	0.04770
Spain	3	0.02539
The Netherlands	17	0.13379
Ukraine	6	0.04953
United Kingdom	5	0.03916
United States	5	0.03944
Total	127	1.00000

N Missing 0
15 Levels

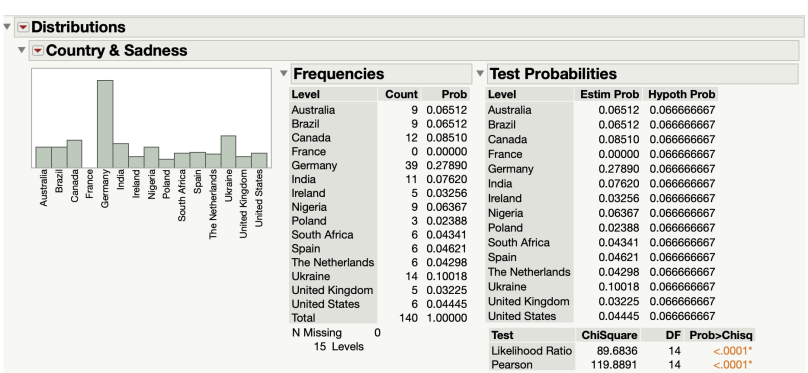
Test Probabilities

Level	Estim Prob	Hypoth Prob
Australia	0.02044	0.066666667
Brazil	0.14309	0.066666667
Canada	0.07013	0.066666667
France	0.13116	0.066666667
Germany	0.04875	0.066666667
India	0.07535	0.066666667
Ireland	0.05366	0.066666667
Nigeria	0.01749	0.066666667
Poland	0.10493	0.066666667
South Africa	0.04770	0.066666667
Spain	0.02539	0.066666667
The Netherlands	0.13379	0.066666667
Ukraine	0.04953	0.066666667
United Kingdom	0.03916	0.066666667
United States	0.03944	0.066666667

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	45.8452	14	<.0001*
Pearson	47.5690	14	<.0001*

of tweets that contain joy. The Chi-square test confirms this, as we receive a p value of 0.0459, lower than 0.05, meaning we reject our null hypothesis. There does appear to be a statistically significant difference among countries in terms of the proportion of tweets that correspond with joy.

For neutral emotions, the graph of distributions shows most of the countries have a varied proportion of tweets that contain neutrality. The Chi-square test confirms this, as we receive a p value of less than 0.0001, less than 0.05, meaning we reject our null hypothesis. There does appear to be a statistically significant difference among countries in terms of the proportion of tweets that correspond with neutrality.



Frequencies

Level	Count	Prob
Australia	9	0.06512
Brazil	9	0.06512
Canada	12	0.08510
France	0	0.00000
Germany	39	0.27890
India	11	0.07620
Ireland	5	0.03256
Nigeria	9	0.06367
Poland	3	0.02388
South Africa	6	0.04341
Spain	6	0.04621
The Netherlands	6	0.04298
Ukraine	14	0.10018
United Kingdom	5	0.03225
United States	6	0.04445
Total	140	1.00000

N Missing 0
15 Levels

Test Probabilities

Level	Estim Prob	Hypoth Prob
Australia	0.06512	0.066666667
Brazil	0.06512	0.066666667
Canada	0.08510	0.066666667
France	0.00000	0.066666667
Germany	0.27890	0.066666667
India	0.07620	0.066666667
Ireland	0.03256	0.066666667
Nigeria	0.06367	0.066666667
Poland	0.02388	0.066666667
South Africa	0.04341	0.066666667
Spain	0.04621	0.066666667
The Netherlands	0.04298	0.066666667
Ukraine	0.10018	0.066666667
United Kingdom	0.03225	0.066666667
United States	0.04445	0.066666667

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	89.6836	14	<.0001*
Pearson	119.8891	14	<.0001*

For sadness, the graph of distributions shows most of the countries have a similar proportion of tweets that contain sadness, with an exception for Germany which has a high proportion of tweets with sadness. The Chi-square test correlates with this, as we receive a p value of less than 0.0001, less than 0.05, meaning we reject our null hypothesis. There does appear to be a statistically significant difference among countries in terms of the proportion of tweets that correspond with sadness.

Conclusion:

Based on our sentiment analysis of the tweets in certain countries, we can conclude that for some emotions, there is an uneven distribution of tweets by country, specifically for the sadness, joy, and neutral emotions. Our hypothesis that Western countries have a large proportion of negative emotions is supported by the data provided. However, what was surprising in our dataset was the abnormal proportion of joy in Ukrainian tweets. We suspect that this is because of small Ukrainian victories, perhaps when the Russians are losing momentum in their invasion. It could also be trying to boost the morale in Ukraine, to help soldiers continue fighting with confidence. Utilizing our sentiment analysis, Ukrainian humanitarian aid initiatives can determine the countries and locations to target their fundraising efforts, especially in places where there is a high proportion of negative sentiment. Political officials looking to best represent public opinion in their foreign policy, such as Russian sanctions, Ukrainian military aid, etc., can also use this analysis to determine their level of commitment and involvement in the war.

Reflection on Process:

In terms of our data collection, we do feel that we could have chosen a better process for determining the sentiment of each country on the Ukrainian War than through their tweets. Only a limited amount of people use Twitter and in addition the tweets that people express online are not entirely representative of their sentiments and the sentiments of a country. In addition, due to the fact that we used existing data and did not use simple random sampling, we are unable to form a conclusion of causation in our observational study.

Works Cited

[1] Purtova, Daria. "Russia-Ukraine War - Tweets Dataset (65 Days)." *Kaggle*, 8 Mar. 2022, https://www.kaggle.com/datasets/foklacu/ukraine-war-tweets-dataset-65-days.

[2] Bellary, Roshan. "Ukrainian-Russian-Conflict-Tweet-Analysis." *GitHub*, 7 Apr. 2022, github.com/roshanbellary/Ukrainian-Russian-Conflict-Tweet-Analysis/blob/main/war.csv. Accessed 25 Apr. 2022.

[3] Garbas, Lukas. "Emotion Classification in Short Messages." *GitHub*, 25 Apr. 2022, github.com/lukasgarbas/nlp-text-emotion/blob/master/traditional_ml.ipynb. Accessed 25 Apr. 2022.