UKRAINE WAR

SENTIMENT AND BOT ANALYSIS REPORT

MAY 2023

PREPARED BY:
JUSTYNA JAKUBOWSKA

Student

PREPARED FOR:
UNIVERSITY OF CALIFORNIA BERKELEY

AI & ML Course Final Assignment

THE PURPOSE

UNDERSTANDING THE PUBLIC SENTIMENT WHILE IGNORING BOTS

CONFLICT SUMMARY

On 24 February 2022, Russia invaded and occupied parts of Ukraine in a major escalation of the Russo-Ukrainian War, which had begun in 2014. The invasion has resulted in tens of thousands of deaths on both sides, and instigated Europe's largest refugee crisis since World War II. About 8 million Ukrainians were displaced within their country by June, and more than 8.2 million had fled the country by May 2023. (Wikipedia)

DATA

I decided to use a dataset that contains almost 48,000 tweets collected in one day in August last year to do a sentiment analysis of the Ukraine war along with the sentiment prediction algorithm. In addition I have used that dataset to design an ML model for predicting if the tweet was posted by a human or a bot.
It is worth keeping in mind that the tweets were in many languages. I chose to concentrate on English tweets and therefore the data decreased to a little bit over 23,000 tweets.

The dataset is from Kaggle and is located here: https://www.kaggle.com/datasets/bwandowando/ukraine-russian-crisis-twitter-dataset-1-2-m-rows?resource=download

PURPOSE OF SENTIMENT ANALYSIS

Sentiment analysis applied to the Ukrainian war can provide valuable insights for multiple stakeholders. For example:

- Government and Policy Makers. Sentiment analysis can reveal public opinion towards the conflict, which can guide policy decisions and strategies. It can also detect shifts in sentiment, possibly indicating changes in the political landscape.
- Humanitarian Organizations. These organizations can use sentiment analysis to understand the feelings of the people affected by the conflict, helping them better target their efforts and communication.
- News Organizations and Journalists. They can use sentiment analysis to gauge public opinion and tailor their reporting accordingly. It can also help identify misinformation or bias in reporting.
- Academics and Researchers. Sentiment analysis can provide a rich dataset for studying the societal and psychological impacts of the conflict. It can also be used to study the propagation of sentiments in social networks.
- Peacekeeping and Mediation Groups. Sentiment analysis can provide insights into the emotional state and attitudes of the parties involved, which could be useful in negotiations and peace talks.

Machine Learning algorithms I explored can be used to automate this sentiment analysis on a large scale. They can process large amounts of tweet data and classify them into sentiment categories:

THE PURPOSE

UNDERSTANDING THE PUBLIC SENTIMENT WHILE IGNORING BOTS

positive, negative, neutral.
After exploring several different approaches I have designed a model that correctly predicts the sentiment of about 85.4% of the tweets.
The potential of sentiment analysis and ML is huge, but one of the key challenges is the complexity and nuance of human language, including sarcasm, idioms, and cultural references as well as native speakers vs. non native speakers.

PURPOSE OF IDENTIFYING BOTS
Creating an ML algorithm to identify bot-generated content is extremely useful in the context of the Ukrainian war. Bots can be used to manipulate public opinion, spread disinformation or propaganda, and exacerbate tensions. Identifying and mitigating these influences has several benefits:

- Disinformation Combat.
  Bots can be used to spread false information at a high speed and scale. Identifying bots can help slow the spread of such disinformation.
- Public Discourse Integrity Maintenance. Bots can artificially inflate certain viewpoints or narratives, distorting the real public opinion. By identifying these bots, we can ensure a more accurate representation of human sentiments.
- National Security.
  Bots can also be used as a tool of information warfare by adversarial nations. Identifying these bots can help protect national security interests.
- Academic Research.
  Understanding the extent and nature of bot activity can provide insights for academic research in fields like media studies, political science, and cybersecurity.
- Social Media Platforms. These platforms have a responsibility to ensure the integrity of their platforms. Identifying bot activity can help them uphold this responsibility.

Creating an ML algorithm to detect bots can be a complex task. It often involves analyzing various aspects of the content and behavior, such as the frequency and timing of posts, the ratio of followers to following, the use of certain keywords or phrases, and the network of interactions.

After exploring several different approaches I have designed a model that correctly classifies users as bot or human approximately 83.84% data.

THE RESULTS

UNDERSTANDING THE SENTIMENT

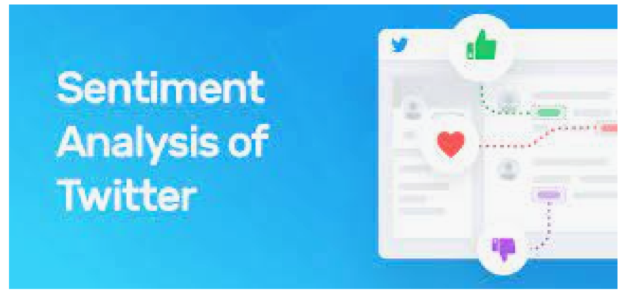
Here are the sentiment results for the original data:

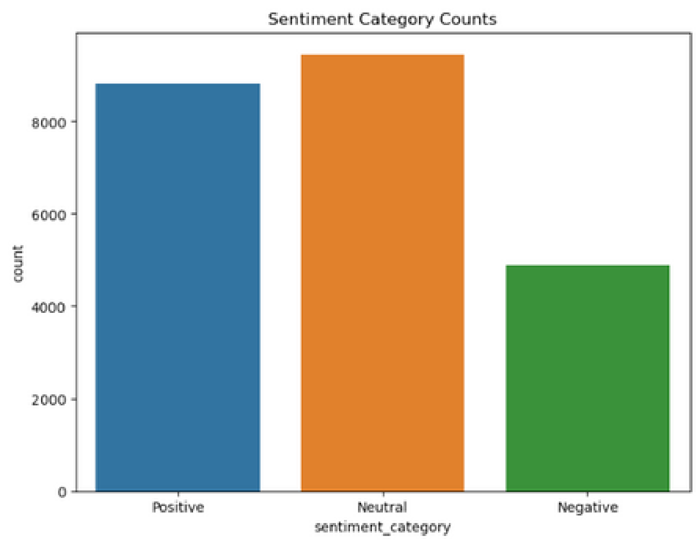
- Neutral sentiment: 9458 tweets (40.804176%)
- Positive sentiment: 8832 tweets (38.103456%)
- Negative sentiment: 4889 tweets (21.092368%)

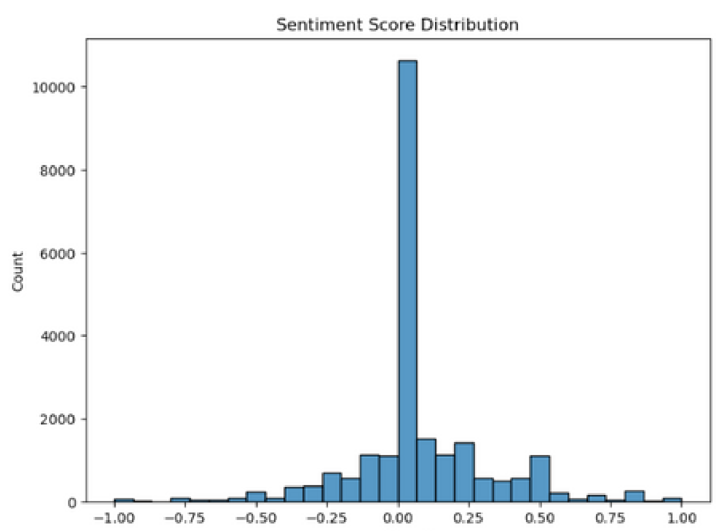
It is best to present the sentiment analysis in a chart form.

In the bar plot "Sentiment Category Counts" on the right, the x-axis represents the sentiment category: Positive, Negative, Neutral and the y-axis represents the count of tweets in each category. The most tweets belong to neutral sentiment category (orange color bar) followed by positive sentiment category (blue color bar) and negative sentiment (green color bar).

In the histogram "Sentiment Score Distribution" on the right, the x-axis represents the sentiment score and the y-axis represents the count of tweets for each sentiment score. The sentiment score ranges from -1 (most negative) to +1 (most positive), with 0 being neutral. The histogram shows how these scores are distributed across the tweets in my dataset.
It looks like overwhelming majority of tweets is somewhere between -0.5 and 0.5 with neutral tweets at 0 leading. That means that there are to many extremely positive tweets (1) or negative tweets (-1).





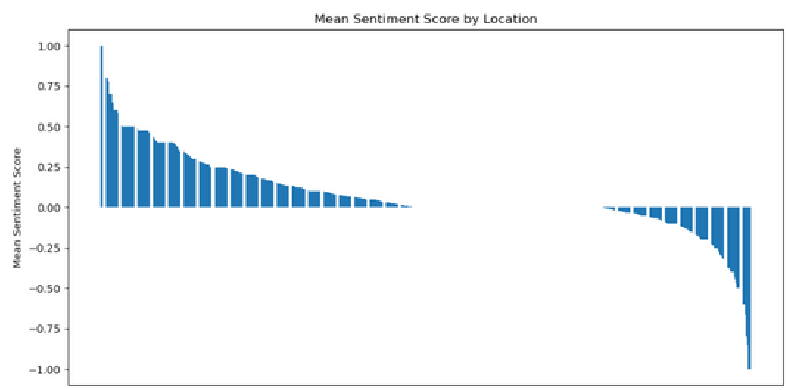


THE RESULTS

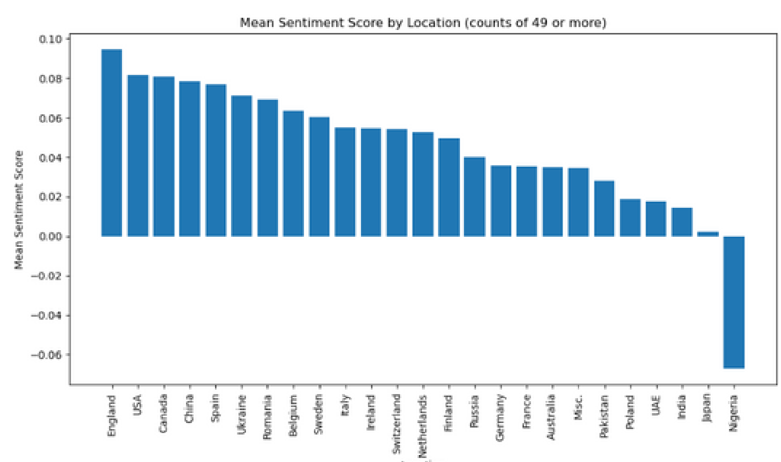
UNDERSTANDING THE SENTIMENT BY LOCATION FOR ALL USERS

SENTIMENT ANALYSIS BY LOCATION
We need to keep in mind that this analysis is more accurate for the parts of the world where English is the native language like USA, England, Canada...
I decided to present the data on 2 charts.

Since there were over 3000 values on x axis I decided not to list them on the chart "Mean Sentiment Score by Location" that is on the left. This chart's goal is to give us a general idea of the sentiment in the world from English speaking population. We clearly can see that majority (about half) of the world has positive sentiment (values above 0). A little bit more than 25% of the world is neutral (value is 0) and slightly less than 25% has a negative sentiment (values below 0). It is also obvious that majority of sentiment is between values of -0.5 and 0.5 and very small population has extreme sentiment about the subject of Ukraine war (values of 1 or -1).



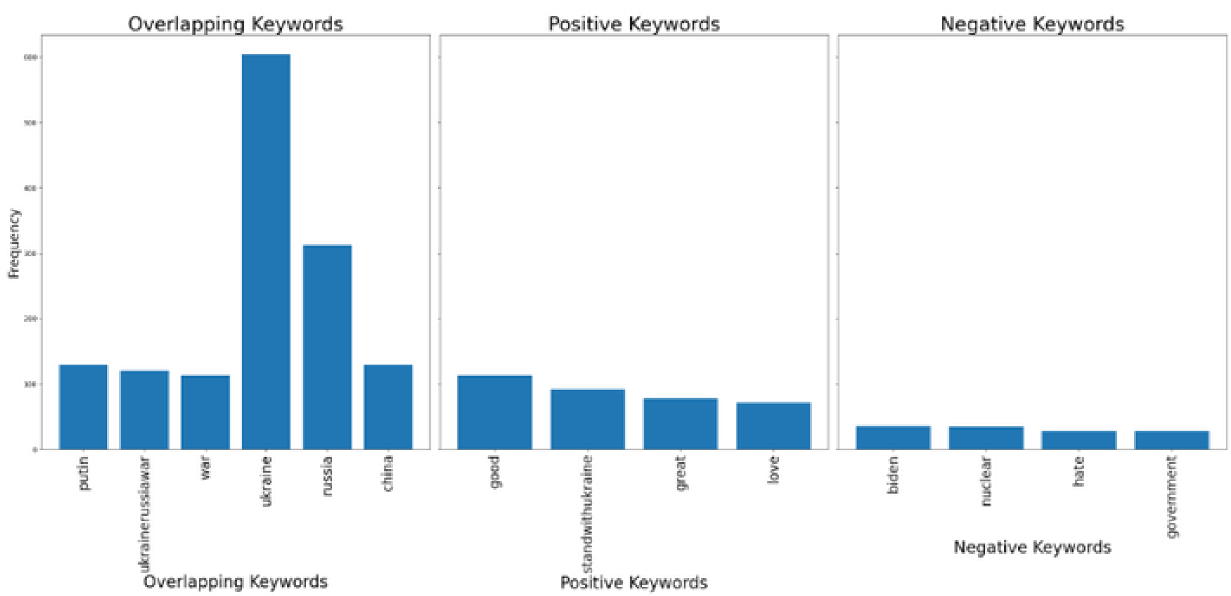
The plot "Mean Sentiment Score by Location (counts if 49 or more)" on the right shows countries' sentiment as well but only for the countries that had 49+ combined frequency. The most positive sentiment shows England that is followed by USA, Canada and China. It is very interseting that the positive sentiment is not very strong and is captured below 0.09 where the maximum scale is 1.0
The most neutral appears Japan that is almost 0.
The most negative is Nigeria, but even then their sentiment is around -0.06 where the maximum scale is -1.



THE RESULTS

UNDERSTANDING THE SENTIMENT BY KEYWORDS FOR ALL USERS

SENTIMENT ANALYSIS BY KEYWORDS RESULTS
The overlapping keywords that repeat in the tweets regardless of their sentiment score are: Ukraine, war, Putin, Russia, ukrainerussianwar and china.
The tweets with positive sentiment have following keywords showing up the most: good, standwithukraine, great, love.
The tweets with negative sentiment have following keywords showing up the most:: Biden, nuclear, hate and government.



THE RESULTS

UNDERSTANDING THE SENTIMENT OVER TIME FOR ALL USERS

SENTIMENT OVERTIME ANALYSIS
I grouped the data by hour. The plot shows how the mean sentiment changes over time in the dataset. It looks like the sentiment mean for the day was positive and fluctuating between 0.03 and 0.09. The sentiment mean was the closest to the neutral state around 11AM and most positive around 3PM.
If we divided data by the time zones we might be able to observe patterns around people daily activities, for example the sentiment can be its highest when people are about to live work since they might be excited about their afternoon plans...