Ukraine invasion Twitter dataset:

About the dataset:

The report delves into an analysis of a dataset centered around tweets concerning Ukraine during the initial month following Russia's invasion. The dataset itself is stored as a 5.4GB CSV file, containing approximately 20 million tweets. The time span of the data spans from February 27, 2022, to March 19, 2022, making it a approximately 1.5 years old.

Dataset fields:

The dataset includes the following fields:

- Userid
- Username
- Account description
- Location of the tweet
- How many other accounts the account is following
- The account's follower count
- The total number of tweets the account has made
- The tweetid
- The timestamp of the tweet
- The retweet count of the tweet
- The text of the tweet
- The hashtags in the tweet
- The language of the tweet
- The coordinates the tweet are tweeted from
- The number of favourites the tweet received
- The extracted timestamp of the tweet

This dataset was selected for its potential to facilitate sentiment analysis related to tweets during the initial month of Ukraine's invasion by Russia.

Dataset collection:

Regarding the data collection process, the dataset itself is hosted on Kaggle. The dataset collector's (Kaggle username: "BwandoWando") utilized Twitter's API to collect the tweets. The dataset was collected using Azure ML with Anaconda notebooks running 24/7 to extract tweets every 15 minutes. The notebooks monitored certain hashtags pertaining to the Ukraine-Russia conflict.

Tests to execute and expected findings:

I have the option to execute the following tests on the data which fall under the following 2 main categories: sentiment analysis of the tweets and bot identification.

Sentiment analysis investigation:

- Get the average sentiment for Russia, Ukraine and other tweet locations
- Get the average sentiment of tweets with high follower counts
- Determine what the most favorited tweets are and their sentiments
- Determine what the most retweeted tweets are and their sentiments
- Create word clouds for Russia, Ukraine and other tweet locations
- Total number of retweets for Positive, Negative and Neutral sentiments
- Create word clouds
- Determine some simple hashtag trends
- Their word clouds
- Do Topic Modelling

Possible bot identification for accounts with:

- Low follower count
- Low following count
- The account user created timestamp, tweet created timestamp, and the extracted timestamp are not similar/close together (only off due to time zone differences)
- If the account has an abnormally large amount of total tweets

The relationships I expect to find are the following: The average sentiment is negative for Ukrainian accounts. The average sentiment for Russian suspected bots will be positive (I expect to find Russian bots that praise the invasion). I will also be able to attempt to: identify bots and see what their sentiment is and identify if the bot's text is misinformation (could be out of the scope of this project).

Investigation into the dataset according to the 5 Vs:

The analysis of the dataset according to the 5 Vs: variety, veracity, volume, velocity, and value. The dataset's variety is significant due to its inclusion of textual data and diverse metadata: geolocation and social media/tweet metrics. The variety of metrics tracked in the metadata are listed above.

The data veracity with respect to the metadata will be high given that smartphones automatically log their data and it is not prone to human error in entering the data. There is however something to be something to be said about human/bot spoofing attempts to

falsify the following data/information: geolocation (with VPNs), timestamps, overall text being false and/or containing misinformation (determining this could be out of the scope of this project) which can lead to altered sentiment of the text data.

The dataset's volume is considerable, since it contains data and metadata from approximately 20 million tweets stored in a 5.4GB CSV file.

Velocity can be viewed with 2 perspectives: one which views the data as high-volume in a short period of time, resulting in high velocity, or the view of low-volume in a long period of time which results in low velocity. I believe that the first perspective is valid, since 20 million related tweets in a month's time is a high volume of data which in a month's time which results in approximately 700 thousand tweets a day. This is approximately 500 tweets a minute or 8 tweets a second, which fits my defenition of high velocity.

Lastly, the value of this dataset lies in its potential to identify public sentiment on the conflict and attempt to identify bot-generated tweets. The bot-generated tweets in the first month of the conflict essentially amount to cyber warfare. Furthermore, bot-driven misinformation campaigns can be shown to impact public sentiment.

The value of the dataset and its analysis can unveil insights into sentiment patterns, identifying potential bots, and revealing trends during the first month following Russia's invasion of Ukraine.

References:

BwandoWando. (2023). (🌇Sunset) 🇺🇦 Ukraine Conflict Twitter Dataset. Kaggle. https://www.kaggle.com/dsv/5934908. DOI: 10.34740/KAGGLE/DSV/5934908.