Manya Sinvhal Kumar

2023 Veritas Fellowship Summer Cohort

Mentor: Mark Lisi

Using Tweets about Climate Change to Determine Public Sentiment

1. **Summary/Abstract:** The purpose of this paper is to view how public sentiment regarding climate change has shifted over time and to identify any trends in sentiment that could prove useful for future climate change and general sentiment analysis. The prediction in terms of the trend of sentiment over time was that the average sentiment would remain slightly negative over time. In this study, the relationship of location, popularity metrics, and emojis to sentiment are also analyzed. The hypothesis for emojis' relationship to the sentiment of a tweet was that a tweet with an emoji is more likely to have positive sentiment. For the locations, the prediction was that even in countries/groups there would not be a general consensus or similar sentiment. The results confirm the hypothesis of a lack of a general trend but, rather than a continuous average negative sentiment, the average remains very neutral over time. In addition, the prediction of a tweet with an emoji tends to have a higher sentiment tweet was also confirmed. To confirm the findings of average sentiment per year, a "weighted popularity" which accounted for likes in addition to the original tweet's sentiment was used. The results regarding location show a distribution of sentiment that doesn't lean one way or another when close together which aligns with the location-related hypothesis. One limitation of this study is lack of editing on the sentiment analyzer. Instead, many

analyzers including VADER and LSTM were used and the most accurate one (VADER) was chosen rather than tweaking and focusing on a singular model which could have made the sentiment labels more accurate.

2. **Introduction:** Over the past few years, researchers have pushed to include sentiment shown in social media to augment the results of polling to best understand public opinion. With so much public data available across many social media platforms, it is important to develop ways to quickly process and understand what the overall sentiments are, without having to manually analyze each post or tweet individually. As this practice becomes more common-place, it will make sentiment analyzers and their addition to polling more effective. This research focused on using recent data from Twitter, analyzing various attributes of tweets like location, use of emojis, and popularity can all help understand more about trends and public perception of climate change.

3. Methods

3.1. Data Sets and Cleaning:

For this project, 3 datasets were used, all of which are publicly available at Kaggle.com. They were processed by using Python, Jupyter Notebooks, and Google Sheets. Each containing various metrics and different pieces of information about the tweets, they contained tweets spanning from the year 2015 to 2022. The three sets are distinguishable by the time periods they cover: 2015 to 2018, 2020, and 2022. All of these are stored as CSV (Comma Separated Values) files. They all contain tweets (string data) and numerical values either representing the popularity of the tweet or the sentiment tag. Only the data set for 2022 contains emoji data. I cleaned the data by removing random characters that appeared during downloading, removing the “RT” that appeared if

it was a response to a tweet (RT signifies re-tweet), and lemmatized the sets. Lemmatization involves simplifying the words so they are easier for the sentiment analyzer to understand. I also did data cleaning by checking for repeat tweets and ensuring the vocabularies were proportional to the number of tweets in each set. The dataset for 2015-2018 had no null values while the data sets for 2020 and 2022 did have null values. The 2020 dataset had null values in the location column which were omitted during the mapping process. A lot of data cleanup was needed because people could manually input their locations. Since there are under 400 tweets in this set, locations set to “Earth”, “Global”, only emojis, and other locations that the geolocater would run into were manually removed. In addition, local spellings of places sometimes differ from how they are spelled in the USA, this made it very difficult to write a program to convert the human made locations into something understandable to the geolocator function so, manually going through each location was the only option. The 2022 dataset has null values mostly under the emoji column when there is no emoji in the tweet; There are also occasional null values in the popularity metrics. None of the datasets have null values that make the data unusable- this would mean not having text in the tweet field. This meant that null values were only removed for a test when there were tests on a column that contained null values. For the 2020 and 2022 datasets, exact times of when the tweet was posted were available. For the 2015-2018 set, only the tweet ID was given. However, by reverse engineering the generation of tweet IDs, I was able to determine the date of the 2015-2018 tweets and get a more precise dataset.

Data Set	Tweet ID	Exact Timestamp	Retweets	Likes	Location	Emojis
2015 to 2018	☑	☑	☐	☐	☐	☐
2020	☑	☑	☑	☑	☑	☐
2022	☐	☑	☑	☑	☐	☑

TABLE 1:Table of known information: *After data clean-up, these were the known aspects of each dataset.*

3.2. Machine Learning

All the tweets were scanned through Python's NLTK pretrained sentiment analyzer: VADER (Valence Aware Dictionary for Sentiment Reasoning). "Topic Modeling and Sentiment Analysis of Global Climate Change Tweets" Dahal et al. (2019) used VADER, making it a reliable option. VADER outputs four values after reading a string of text: how positive it is, how neutral it is, how negative it is, and a compound value that ranges from -1 to 1 which shows what VADER thinks the overall sentiment is. To test its accuracy, the VADER results were compared to the polarity labels that were in the 2020 set. However, since the 2020 polarity ranges from -0.5 to 0.6 and the VADER compound sentiment label ranges from -1 to 1, polarity times 2 was compared to VADER. To check accuracy, I used MSE (mean square error) by subtracting VADER's prediction from polarity and summing the squares of each of these values. In the 2020 dataset, the MSE was 0.1878. The sentiment was overall positive which was the same as the polarization VADER was being compared to making it the sentiment analyzer for the entire study and was applied to the rest of the datasets.

4. **Results:** To find the overall sentiment, sentiment given by VADER was averaged in each data set. After the exact timestamps for the tweets into the 2015-2018 dataset were found they were split by year. Below are the average sentiments per year in graph and table form:

2015	2016	2017	2018	2020	2022
0.0034755	-0.037218	-0.033134	-0.003272	0.000474	-0.00196

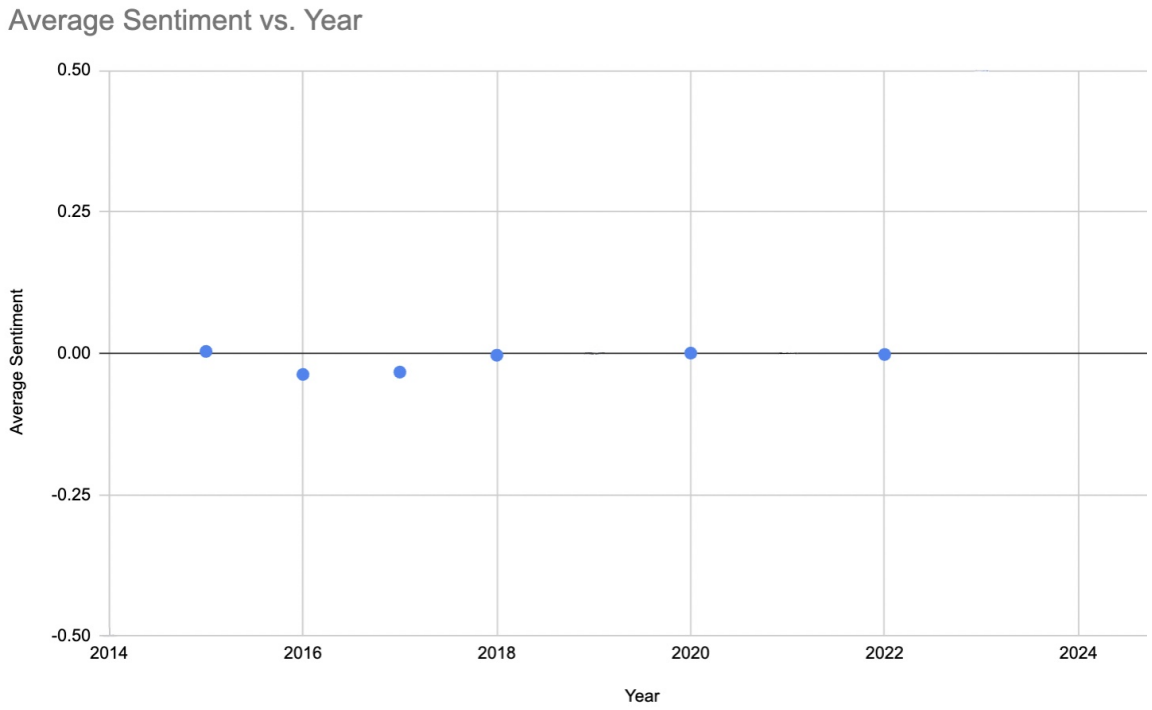

TABLE 2 and FIGURE 1: 2015 to 2022 Tweets Plotted by Sentiment: *The timestamps were either given by the datasets or obtained by reverse engineering the tweet ID. The sentiment is on a scale from -1 to +1 given by VADER. The more negative a sentiment is, the more negative the overall mood of the tweet is, the more positive it is, the opposite is true.*

The geo-tagged tweets were plotted on a map based on longitude and latitude to search for any general trends between sentiment and location:

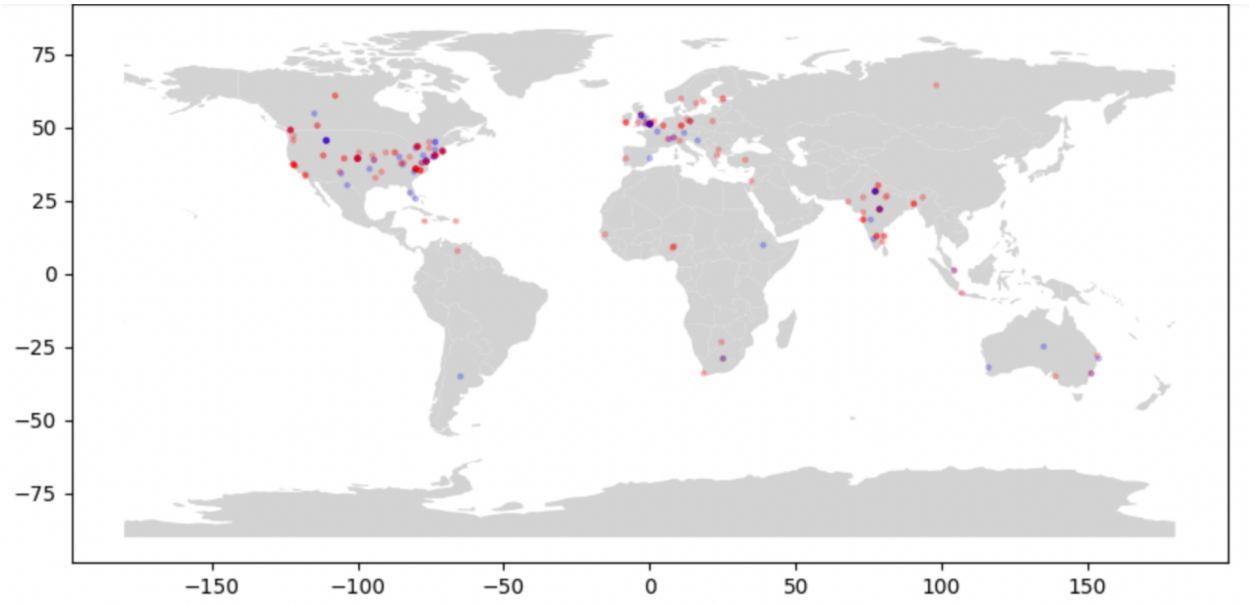

FIGURE 2: 2020 Tweets Plotted by Location and Sentiment: *Blue dots correspond to negative sentiment and red dots correspond to positive sentiment. Each point has an opacity of 20%. Darker dots mean more tweets in the same region.*

No general trend was found connecting sentiment to location.

The 2020 and 2022 datasets included various popularity metrics including comments, likes, retweets, and friends of the user. This research used an algorithm to increase the weightage of a tweet by factoring in the number of likes; Only likes were considered in "weighted popularity" because it was the metric that has the same sentiment as the original tweet. This weighting algorithm did not consider the number of comments, friends, or retweets since they are complex in terms of agreeing or disagreeing with the original tweet. The "weighted popularity" of a tweet was calculated by raising the number of likes to the 1.01 power and then multiplying it to the sentiment given by VADER.

[*"weighted popularity" = (Number of Likes^1.01) * sentiment]*

Unweighted 2022 Sentiment	-0.00196
Weighted Popularity 2022 Sentiment	-1.406
Unweighted 2020 Sentiment	0.000474

Weighted Popularity 2020 Sentiment	1.730

TABLE 3: Weighted vs. Unweighted Sentiment: *The number of likes in the 2020 and 2022 datasets were used to calculate a weighted sentiment to confirm the average sentiment. This was to ensure the average sentiment and the tweets that would get the most support were either both positive (2020) or both negative (2022).*

This test was to see if the average sentiment of tweets was the same with and without these calculations that accounted for the popularity of each tweet. The average 2022 sentiment with no account of popularity was -0.00196. The "weighted popularity" was -1.406. Since the popularity of the tweets remained negative, it reinforces the finding that the average 2022 sentiment was indeed negative. For the 2020 sentiment, the weighted sentiment confirmed that the sentiment was indeed positive because the positive tweets were the majority and got the most likes.

One unique aspect of this (paper/study) is its look at emojis- typically, published papers omit emojis. One important aspect to note about this test is that emojis were omitted from the text analyzed by VADER- they were stored in a separate column. The aim of this test is to determine if emoji use or a certain emoji can serve some indication of whether the tweet had a positive or negative sentiment. This section will only be looking at the tweets of 2022 and comparing it to the overall tweets for 2022 because this is the only dataset with emojis. The average sentiment of all tweets in 2022 was -0.001963303867. Only looking at the sentiments of tweets that contain emojis the average was 0.03120150943.

Average 2022 Sentiment	-0.00196
Average 2022 Sentiment of only tweets that included emojis	0.0312

TABLE 4: Emoji vs. No Emoji Sentiment

This demonstrates the use of emojis suggests a more positive sentiment.

5. **Discussion:** Researchers have looked at tweets surrounding climate change to understand best how location, gender, and climate events affect sentiment on environmental issues. "Topic Modeling and Sentiment Analysis of Global Climate Change Tweets" Dahal et al. (2019) looked at over 300,000 tweets across the globe using geospatial and temporal perspectives. They found an overall negative (defined as an unhappy emotional state) reaction to climate issues that were in direct response to real-life events or influential tweets such as one by former President Donald Trump. In "A Demographic Analysis of Online Sentiment during Hurricane Irene" by Mandel et al. (2012) looked at tweets following the disastrous Hurricane Irene by viewing tweets from impacted areas, analyzing how gender affected sentiment and levels of concern leading up to the hurricane. The results demonstrated that a twitter user's gender and proximity to a natural disaster were related to sentiment. Although there were 44,000 tweets used in this paper, it is important to note that it is only a small fraction of all climate change related tweets. In addition, with the current change of leadership in Twitter, its rebranding, and new rules of use, Twitter may not be a viable option in the future for researchers to get accurate and all of the public's data. As of February 9th 2023, Twitter policies have eliminated free APIs and created paid subscriptions up to $210,000 per month for accurate and large amounts of data. However, this research is not exclusive to Twitter and can be used and applied to other posts such as Meta's new social media app: Threads.
6. **Conclusion:** The findings show that sentiment regarding climate change from the years 2015 to 2022 on Twitter averages to be neutral. This does not mean that the majority of sentiments are neutral but rather there is a strong split in opposing views. These findings support my hypotheses which show the importance of using social media as a way to help

in properly finding the correct sentiment of the public. Since VADER is a general sentiment model and all cleaning could be applied to any dataset this can easily be generalized by applying it to data in fields other than environmental science. Although social media should not be used as a substitute for polling, it can be seen that social media allows researchers to analyze opinions users post.

7. References:

https://www.researchgate.net/profile/Sathish-Kumar-26/publication/331453828_Spatiotemporal_Topic_Modeling_and_Sentiment_Analysis_of_Global_Climate_Change_Tweets/links/5d9033e6a6fdcc2554a4740e/Spatiotemporal-Topic-Modeling-and-Sentiment-Analysis-of-Global-Climate-Change-Tweets.pdf

https://aclanthology.org/W12-2104.pdf

https://ojs.aaai.org/index.php/ICWSM/article/download/22194/21973

https://www.washingtonpost.com/technology/2023/06/20/twitter-policy-elon-musk-api/