CSC110 Project Report: Climate Change Sentiment on Twitter

Harsh Jaluka, Ronit Kumar, Thomas Liu, Wilson Sy

December 14, 2020

Problem Description and Research Question

Evidence of climate change and its potential harms date as far back as the early 19th century. Since then, after decades of research, the vast majority of scientists have reached the conclusion that the increase in global temperature is, in large part, due to human activities; particularly, the increase in greenhouse gas emissions after the industrial revolution.

Despite the evidence for anthropogenic global warming, or man-made climate change, there is a lack of agreement about the very existence of climate change among the general public. There is a vocal group of people around the world that denounces the legitimacy or the severity of climate change. These people may be motivated by political or economic gain, or they may themselves be victims of misinformation disseminated by social media.

Research shows that polarization can be attributed to influencers, whose views are often amplified over others in a network (Centola, 2020). In political discourse on Twitter, such influencers can be major politicians: a study about politicians in the 2016 U.S. election found that tweets with "more emotive and moral words" get retweeted more often (Brick, Linden, & De-Wit, 2019). Thus, in a race to gain more popularity, politicians often intensify their rhetoric, whether they believe or disbelieve in climate change. Moreover, Ezra Klein from Vox Media suggests that politicians' dependence on social media to read the will of the people may polarize politicians' opinions, leading to more unsatisfying choices when the public is faced with polarizing public policy that aims to tackle climate change (Newton, 2020).

Climate change is a serious, pressing issue. Consensus among the public about its legitimacy is vital to allow for concrete action against it. In our project, we want to study the intensity of the rhetoric between opinions for and against climate change by doing sentiment analysis on tweets. We are hoping to determine the extent to which the debate surrounding climate change has been polarized, primarily by comparing opinion-based tweets with neutral fact-based tweets. Thus, our research question is as follows:

How polarized is the debate surrounding climate change on Twitter?

Dataset

We collected information from the Twitter Climate Change Sentiment Dataset by Edward Qian (2019).

The dataset is stored as a `.csv` file, where each row contains the relevant data for a tweet. There are three columns: the sentiment value for the tweet, the content of the tweet, and the id of the tweet. (In our code and from here on in the report, we refer to this sentiment value as 'opinion value' to avoid confusion with values calculated by our sentiment analysis library.) Only the opinion and the content columns of the dataset are relevant for this project. The opinion is one of four values:

-1 : The contents of the tweet do not support the idea of man-made climate change

0 : The contents of the tweet are neutral on idea of man-made climate change.

1 : The contents of the tweet support the idea of man-made-climate change.

2 : The contents of the tweet link to factual news on climate change.

The dataset has 43 943 tweets on climate change, collected from April 27, 2015 to February 21, 2018. The tweets' opinion values were assigned by 3 independent reviewers; only tweets where all 3 reviewers agreed on a value were kept. In the dataset, there are 3990 tweets (9.0%) that do not support man-made climate change, 7715 neutral tweets (18.0%), 9276 tweets about news on climate change (21.0%), and 22 962 tweets that support the idea of man-made climate change (52.0%).

Our program stores the opinion value and the content of each tweet as instance attributes of the `Tweet` dataclass.

Computational Overview

New Library

`vaderSentiment` is the new library at the centre of our project. It is a sentiment analysis library that calculates the polarity (Is the writer happy or angry?) and intensity (Is the writer mildly upset or outraged?) of any given text. It calculates the proportion of words and symbols in a text that express a particular emotion: negative, neutral, or positive. Then, using these three values, it calculates a compound score of the overall sentiment of the text, ranging from -1 for extremely negative to 1 for extremely positive (Hutto, 2020).

Data Transformation

Our program uses instances of the `Tweet` dataclass to represent each tweet. The class has three instance attributes: `opinion`, `content` and `sentiment`. The first two correspond to columns in the raw dataset. The last attribute `sentiment` stores the polarity scores of each tweet (as calculated by the `vaderSentiment` library). It is an optional attribute and set to `None` initially. The `process` function takes in the filepath of the dataset, reads the corresponding file, and outputs a `List` of instances of `Tweet`, each representing an individual tweet. The `sort_tweets` function takes in this list of tweets and outputs a dictionary that categorizes tweets by their `opinion` attribute.

We encountered some difficulties with the dataset because some characters in the content of the tweets were decoded incorrectly as ISO-8859-1 and not Windows-1252. This created weird characters in our stored text, which `vaderSentiment` was unable to analyze. Our implementation of `process` attempts to fix this issue by encoding the text as Windows-1252 and then decoding as UTF-8 (Tex Texin, 2011). The workaround fixes most of the encoding errors; however, some characters remained unfixed for reasons that we could not identify. Fortunately, only punctuation and special characters are affected by this lingering error; leaving these unfixed would not make a significant difference to the sentiment analysis.

We also filtered out some substrings within each tweet's content. In particular, we removed all hyperlinks, mentions and hashtags because these would be falsely identified as neutral strings by the `vaderSentiment`'s lexicon (as these words do not exist in the lexicon) (Hutto, 2020).

VADER Analysis

Using the aforementioned `vaderSentiment` library, our program finds the polarity scores and compound score of each tweet. Specifically, our program uses the `polarity_scores` method of an instance of the `SentimentIntensityAnalyzer` class in `vaderSentiment`. The method outputs a dictionary with four keys, `'neg', 'neu', 'pos', 'compound'` and the value of each key is the value calculated by the method based on the content of the string. Each Tweet instance's `sentiment` attribute was then updated accordingly by the `add_vader_to_tweets` function.

Displaying Results

Using `plotly`, our program plots multiple graphs that can be interactively accessed in a `pygame` application.

To obtain a crude understanding of the data, `compare_frequency_vader` displays a grouped bar chart. The x-axis of the plot represents the four opinions on climate change; for each opinion, there are three bars, one for each type of tweet (positive, negative and neutral). The tweets are classified based on the compound score calculated by the `polarity_scores` method. A positive tweet is one that has a compound score between 0.05 and 1 inclusive, a negative tweet has a compound score between -0.05 and -1 inclusive and a neutral tweet is one that has a compound score between -0.05 and 0.05. Each bar represents the percentage of tweets of that sentiment within the given opinion.

`normal_histogram` is a function that takes a list of Tweets and creates a Figure object in `plotly` that displays a normalized histogram. The histogram displays the distribution of compound scores (which can be accessed through a Tweet's `sentiment` attribute using the `'compound'` key). The `plotly` graph also displays some summary statistics in an annotation below the graph: these are calculated by `summary`.

`normal_histogram` is called on lists of tweets that have been sorted by opinion value (using `sort_tweets` to show the distribution of the compound scores for each climate change opinion in our `pygame` application.

`plot_pos_neg` takes in tweets sorted by opinion value and creates a Figure object in `plotly` that displays a scatter plot of tweets. The scatter plot's x-axis shows the negative scores, and the y-axis shows the positive scores;

both values are accessed through the `sentiment` attribute of Tweet.

`plot_compound` takes in tweets sorted by opinion values and creates a Figure object with multiple boxplots. The boxplots are of the value associated with the 'compound' key of the dictionary sentiment attribute from the Tweet class. Then we use compound scores of the tweets in lists sorted by opinion value to create boxplots that can be used to compare the distribution of these compound scores to each list.

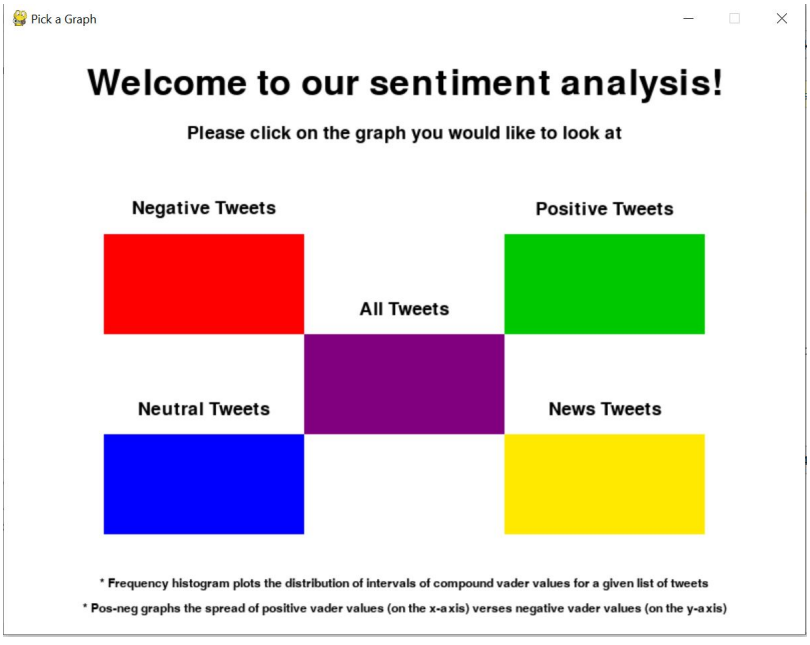
The `pick_graph` module contains a `pygame` application that acts as a menu to access graphs. There are four boxes in each corner with a subtitle above them indicating the type of tweet the graph corresponds to. The user can hover over one of the displayed boxes and choose the histogram of compound scores produced by `normal_histogram`, or the scatter plot of negative and positive scores of each tweet. When either option is clicked, the corresponding graph will be displayed in a new window on your browser. In the center, there is a box for all tweets, where the user can hover over to either see the box plots of compound scores produced by `plot_compound`, or the grouped bar chart produced by `compare_frequency_vader`.

Running the Program

1. Install required libraries listed in `requirements.txt`.

2. Download the dataset from this Kaggle page. (Alternatively, pick up the file from UTSend using Claim ID `HwvZms4iQHJxT3bw`, and passcode `EexmQcMcc7D9CdoP`.) Place the `.csv` file in the root directory of the project folder; at the same level as `main.py`.

3. Run `main.py`. A `pygame` window should open (as shown in the image below), displaying multiple options of graphs of the data. Hovering over each box will reveal two options to choose from for each set of Tweets. Upon selecting an option, the corresponding graph will open in the user's default internet browser.



Description of Changes

Each tweet was originally supposed to be stored as a size-3 tuple, containing the opinion, content, and id. The first change here was that we decided to use a dataclass instead of a list as it would be more convenient to call on each attribute of the tweet in other parts of the project. The second change was that we stopped storing the `id` column as it was not of any use in the project. The third and final change to the tweet data structure was the addition of an instance attribute called `sentiment` that would store the polarity scores of each tweet.

According to the documentation of the `vaderSentiment` library, a neutral sentiment is exemplified by a compound value between -0.05 and 0.05. However, instead of taking `vaderSentiment`'s suggested range, we had initially planned to calculate our own range for neutral sentiment by finding the range of compound values for tweets which have the opinion 'news' or 'neutral'. When this range was computed, we found that the range was actually quite large, a contrast from what we were expecting. It turned out that our hypothesis that non-neutral non-news tweets having more extreme compound values was false. It was then that decided that we simply use the suggested range for neutral sentiment as the suggested range is a standard that would allow us to gauge just how polarized tweets surrounding climate change are.

Although `vaderSentiment` was essential to the analysis that we conducted, there was not much that we could use it for apart from calculating the polarity scores of the content of each tweet; we did not use it as 'extensively' as may have been expected of us. However, we realized that in order to interpret the data from the `vaderSentiment` library, we had to plot graphs that would shed light on our research question. We also had to decide on a way to display the graphs. Hence, we decided to incorporate extensive use of `plotly` and `pygame` in our project. `plotly` was used in visualizing all of the data from `vaderSentiment` and, as discussed above, `pygame` was used to create an application that would act as a menu to access the graphs.

Besides changes to our implementations and data analysis, we changed some wording to reduce human error in writing the code for our program and to reduce confusion. The most significant of these changes was to rename the 'sentiment' value extracted from the dataset as 'opinion'. Prior to the change, there was confusion as to whether 'sentiment' should refer to the dataset column or the scores provided by VADER analysis.

Discussion

Results

Under the 'All Tweets' section of the `pygame` application, the graph obtained by clicking 'Opinion' visualizes the first statistical analysis that we conducted on the information from the `vaderSentiment` library. For each opinion, it shows the percentage of tweets that fall under each sentiment category (positive, neutral or negative). Looking at the tweets that do not support climate change, some key takeaways are that 47% of these tweets were classified as negative and 38% as positive. This indicates that the rhetoric among those who do not support climate change

does seem to be quite harsh, especially when compared to those with other opinions. Indeed, taking a look at the tweets that do support climate change, we see that the percentage of positive tweets remains roughly the same and the major increase in the percentage of neutral tweets comes from a sharp decrease in negative tweets. Now, taking a look at the remaining two categories, which we had initially hypothesised to be more neutral, we see that our hypothesis was only partially correct. While these two categories do have the highest percentage of neutral tweets, this percentage only went up as high as 36% for the news tweets. Surprisingly, 48% of the neutral tweets were classified as positive (the highest) compared to only 28% of the news tweets (the lowest). Evidently, the rhetoric of news and neutral tweets are quite in contrast with each other.

The 'Sentiment' graph under 'All Tweets' displays a box plot for each opinion. One of the first things that can be noticed is that the interquartile range for the supporting and not supporting tweets is significantly greater than the IQR for neutral or news tweets. Since the median is zero for all four sets of tweets, this is not surprising as the news and neutral opinions had the highest percentages of neutral tweets, which would lead to more tweets with a compound value near zero. Some of the key takeaways from this graph are that the first quartile (Q1) is -0.44 for tweets from the 'do not support' camp and -0.36 for tweets that do support man-made climate change. This implies that tweets under the category 'do not support' have more extreme negative compound scores compared to the other categories. As a result, the tweets that do not support climate change skew the most negatively. Another interesting conclusion that we can draw is that the neutral tweets appears to show the most positive sentiment as both its first quartile (Q1) and third quartile (Q3) have greater compound values than all other opinion categories. In fact, the box plot that skews the most positively is the neutral tweets, while the tweets in support of climate change is the most balanced.

The 'Frequency Histogram' graphs for each opinion have a huge spike at 0. This could be because of two reasons; either a large percentage of tweets do not contain many words with high polarity scores, or a large percentage of tweets contain a lot of words that are not in the `vaderSentiment` lexicon. The actual reason is likely a combination of both and we believe that a lot of words not being in the lexicon is the major contributor. The lexicon consists of 7,520 strings (Hutto, 2020) that `vaderSentiment` calculates the polarity scores based on while the dataset had 43,943 tweets. There certainly are a lot of words within the dataset that do not exist in the lexicon. Tweets are also not spelled entirely correctly; a lot of words may have been falsely identified as neutral due to a spelling error or other spelling variations.

The 'Pos-Neg Graph' for each opinion also supports the conclusion reached from the 'Frequency Histogram' graphs. Each graph shows a concentration in the bottom left, as both positive and negative polarity scores go to zero. Since the positive, negative and neutral polarity scores represent the percentage of text that has that sentiment within the string (Pandey, 2018), a concentration around zero for both positive and negative polarity score implies that a large number of tweets have a large neutral polarity score. This explains the spike at zero for each of the histograms. We plotted this scatter plot with the intention to look for any significant skewing. There are some

aspects which affirm what we have seen in previous graphs. The scatter plots of the tweets that support climate change, is much more concentrated around lower negative scores that have greater positive scores. Moreover, news-based tweets and tweets that do not support climate change have markedly less positive scores, so much so that `plotly` has a smaller y-axis scale for the two. However, though the data clusters near the origin, the graphs mainly show that the due to the discrepancy in the number of tweets of each opinion and the relatively normal distributions shown in the histograms, we could not reach any major conclusion with these scatter plots.

Further Exploration

One of the major limitations of our analysis was the breadth of the VADER lexicon and the unpredictable content of tweets. Despite our efforts to filter the content of the tweets by removing problematic substrings like hyperlinks, mentions, and hashtags, a lot of other errors, which are not as easy to account for (spelling errors, in particular) slipped past our filters. There is room for further exploration in methods to account for these errors and to get more accurate compound values for each tweet.

Another avenue of further exploration could be to use the Twitter API to collect real-time data and plot some graphs of overall sentiments and opinions of climate change over time. Our current project only looked at tweets from a chosen time period. We believe current data about the climate change may show that discussion on social media may exhibit greater signs of polarization; particularly as it has become a more politically prominent issue.

Conclusion

Analysis on the dataset does not point to a high degree of polarization in climate change discourse on Twitter. The sentiment distributions for opinions supporting and denying climate change are very similar: the percentage of positive, neutral, and negative sentiments expressed over each opinion are close, as seen in the Opinion chart of all tweets. Both distributions' medians are 0, and their means are both slightly negative. However, due to sources of error in our analysis, such as unanalyzable characters in tweet contents, the exact extent of non-polarization is somewhat unclear.

Looking at subsets of the data, we can see some trends that may suggest polarization. For example, tweets denying man-made climate change have the most negative-skewed composite scores, while tweets in support of climate change are largely balanced around the median. Such a difference can be explained by the contrasting perspectives of hope for change and a fear of the dire consequences of inaction. Again, it is inconclusive whether this skewing is significant.

The news tweets skew negative considerably as well and have the lowest Q3 of the opinion categories; this may be explained by less reliable news outlets presenting information in a negative light to garner more attention (and revenue). The most surprising finding was how strongly the neutral tweets skew positive; almost 50% of neutral

opinion tweets were measured to have positive sentiment. We believe the findings on neutral tweets warrant closer examination. Are these tweeters just in blissful unawareness of climate change? Or perhaps an approach to tackling climate change less involved with eliciting emotional reactions may unite action against climate change and evoke more hope for a better future.

References

Bar Charts in Python. Plotly. https://plotly.com/python/bar-charts/

Box Plots in Python. Plotly. https://plotly.com/python/box-plots/.

Brick, C., Linden, S. V. D., & De-Wit, L. (2019, January 16). *Are Social Media Driving Political Polarization?* https://greatergood.berkeley.edu/article/item/is_social_media_driving_political_polarization

Burchell, J. (2017, April 8). *Using VADER to handle sentiment analysis with social media text.* https://t-redactyl.io/blog/2017/04/using-vader-to-handle-sentiment-analysis-with-social-media-text.html

Centola, D. (2020, October 15). *Why Social Media Makes Us More Polarized and How to Fix It.* Scientific American. https://www.scientificamerican.com/article/why-social-media-makes-us-more-polarized-and-how-to-fix-it/

Cook, J., Nuccitelli, N., Green, S. A., et al. (2013, 15 May). Quantifying the consensus on anthropogenic global warming in the scientific literature. *Environmental Research Letters*, 8(2). https://iopscience.iop.org/article/10.1088/1748-9326/8/2/024024

Histograms in Python. Plotly. https://plotly.com/python/histograms/.

Hutto, C. J. (2020). *VADER-Sentiment-Analysis.* https://github.com/cjhutto/vaderSentiment

Malde, R. (2020, June 8). *A Short Introduction to VADER.* https://towardsdatascience.com/an-short-introduction-to-vader-3f3860208d53

Maslin, M. (2019, 30 November). Here Are Five of The Main Reasons People Continue to Deny Climate Change. *Science Alert.* https://www.sciencealert.com/the-five-corrupt-pillars-of-climate-change-denial

Misri, A. (2020, 8 May). How to create Buttons in a game using Pygame? *GeeksforGeeks.* https://www.geeksforgeeks.org/how-to-create-buttons-in-a-game-using-pygame/

NASA. (2020). *Scientific Consensus: Earth's Climate is Warming.* Global Climate Change: Vital Signs of the Planet. https://climate.nasa.gov/scientific-consensus/

Newton, C. (2020, February 28). Why we can't blame social networks for our polarized politics. *The Interface.* https://www.theverge.com/interface/2020/2/28/21153060/social-network-polarization-ezra-klein-why-were-polarized-q-a

Pandey, P. (2018, September 23). *Simplifying Sentiment Analysis using VADER in Python (on Social Media Text).* Medium. https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f

plotly.graph_objects.Bar. Plotly. https://plotly.com/python-api-reference/generated/plotly.graph_objects.Bar.html.

Python Figure Reference: layout.annotations. Plotly. https://plotly.com/python/reference/layout/annotations/.

Python: How to read and write CSV files. thepythonguru.com. (2020, January 7). https://thepythonguru.com/python-how-to-read-and-write-csv-files/.

Qian, E. (2019). *Twitter Climate Change Sentiment Dataset.* https://www.kaggle.com/edqian/twitter-climate-change-sentiment-dataset

Scatter Plots in Python. Plotly. https://plotly.com/python/line-and-scatter/

Sentdex. (2014, 10 October). Pygame Buttons, part 1, drawing the rectangle. *Python Programming.* https://pythonprogramming.net/pygame-buttons-part-1-button-rectangle/

Tex Texin. (2011). *UTF-8 Encoding Debugging Chart.* https://www.i18nqa.com/debug/utf8-debug.html

Weart, S. (2017, August 17). *The Discovery of Global Warming [Excerpt].* Scientific American. https://www.scientificamerican.com/article/discovery-of-global-warming/