Climate-Change-Sentiment-Predictor

This project aims to develop a program that, given a tweet relating to climate change, can predict whether the user's sentiment is sympathetic regarding climate change or not. Using data from a dataset of tweets with a manually labeled sentiment to train from, we aim to help future analysis regarding climate change through this program that looks into human sentiments through tweets. By providing a program that automatically labels sentiments, this project can easily provide future research with numerous and timely data to use.

This project is developed by:

- Alvarado, Enrique Luis
- Bassig, Lance Raphael
- Roxas, Dennis Christian
- Surara, Ron Christian

Definition of Terms

1. **Sympathetic**: *Adjective describing someone who talks and acts for a particular topic/sentiment: in this study, someone who is aware of, and/or acts in a way that makes others be positively aware of, climate change.*
2. **"Positive"**: *Label of climate change tweets that attempt to bring positive awareness to climate change and their causes through information or persuation. In essence, this label is attached to tweets aligned with the sentiment that humans are the cause of climate change and that it is an urgent issue. This label also includes news on climate change.*

   > Indiscriminately bang nilabel ang true and fake news under this label?

3. **"Negative"**: *Label of climate change tweets that negatively bring awareness to climate change and their causes, e.g. deniers of climate change or stating that the issue is out of our control thus dissuading action.*

   > Ininclude ba ang fake news sa label na ito?

4. **"Neutral"**: *Label of climate change tweets that simply express awareness of the issue but not attempt to inform, persuade nor dissuade people from it are labeled "neutral".*

Problem Statement

This project aims to lay the groundwork towards answering the following question: "How sympathetic can we expect the general populace to be with regards to climate change?"

In particular, the research group seeks to do the following:

1. Perform Exploratory Data Analysis (EDA) on the composition and word occurences of the data set from which the machine will learn from,
2. Choose the best Logistic Regression model to use,
3. Utilize the learned machine to label climate change related tweets from May 2016 to May 2017, and from May 2021 to May 2022.
4. Perform EDA on the composition and word occurences of the resulting labeled data, and

5. Determine if there is a significant change in sentiment on climate change over the past half decade.

The null hypothesis is determined to be that "there is no significant change in the twitter sentiment on climate change from 2016-2017 and that of 2021-2022."

The alternative hypothesis is determined to be that "there is a significant change in the twitter sentiment on climate change from 2016-2017 and that of 2021-2022."

Methodology

The machine that we have developd was be made to learn how to classify climate-change related tweets to the following categories: "positive", "negative", and "neutral". (See the definition of terms for what each of the labels mean)

Each tweet in the learning data that was used by the machine was individually labeled by the research group according to their stance and sentiment on climate change; the Climate Sentiment on Twitter (Guzman, 2020) dataset was reviewed for this purpose. This dataset is a raw database of 396 tweets from January 2020 to December 2020. The data from this dataset is just large and recent enough to be utilized for this project.

The data was preprocessed by adding a new `Sympathy?` column that indicates whether a tweet is sympathetic to climate change or not; then the group manually added the appropriate value for each tweet ("Yes" if the tweet is sympathetic to climate change, "No" otherwise). Instances of null rows were removed, and `TfidfVectorizer` was then used to convert text data to numeric data. The tweet content will be cleaned of its URLs, hashtags, mentions, emojis, smileys, and stop words (e.g. "a", "the", "this" , "amp", which is the HTML code of the ampersand symbol, `&amp`); this is so that the machine can better process and properly learn the necessary information from the dataset. Exploratory data analyses on the sentiment composition and word occurences were also performed on this dataset to have an idea of what the machine learns from the tweets.

The Logistic Regression model, particularly the Limited-Memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) solver, will be used in this project. According to a prior analysis (Kim, 2022) on a similar dataset (Qian, 2019), this model has been observed to perform best particularly in distinguishing tweets on climate change to appropriate sentiment labels.

After the machine has been developed, tweets related to climate change was scalped from twitter and fed to the machine. The resulting labeled data was then explored in order to answer the main problem faced by this project.

Specifically, 1059 tweets related to climate change between May 2016 to May 2017 and 1002 tweets related to climate change between May 2021 to May 2022 were then scalped from twitter for the machine to label. Exploratory data analysis is performed on the resulting labeled data to have an idea on what the sentiments of the tweets are like.

Finally, a chi-squared test is performed on the sentiment counts of the results of the two datasets labeled by the machine to determine whether to reject the null hypothesis or not, with `alpha = 1 - model accuracy`.

Results and Discussion

The composition of the dataset used to train the machine can be observed in the following graph:

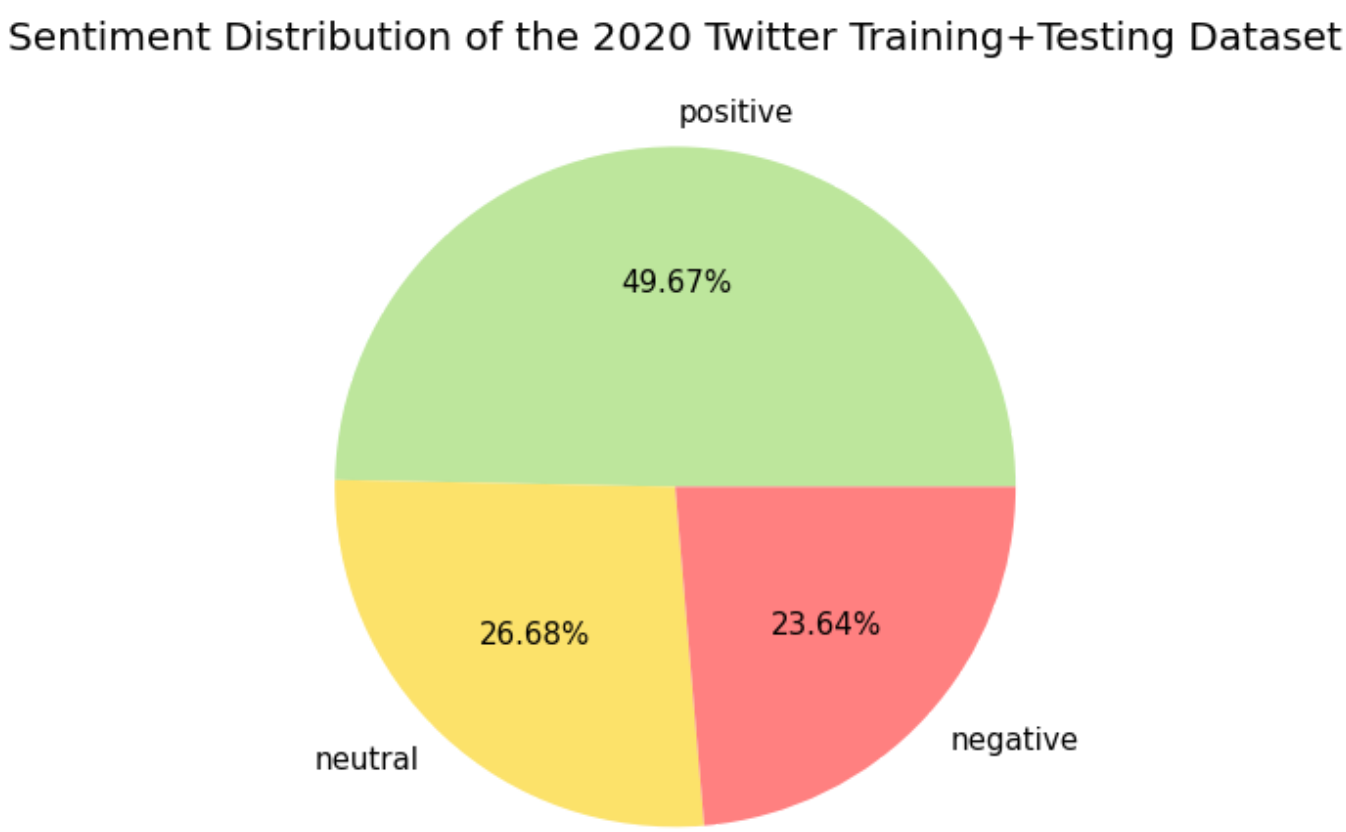

Figure 1. Sentiment Composition of the dataset used to train the machine.

Almost half of the training dataset is composed of positive tweets, while the remaining half is composed of a roughly equal number of neutral and negative tweets, with negative tweets having the least number in the dataset.

When the dataset was cleaned and tokenized, the word occurrences and compositions of tweets of respective sentiments can then be determined.

The following bar graphs show the occurrences of the top 30 most occuring words among tweets of the respective sentiment:

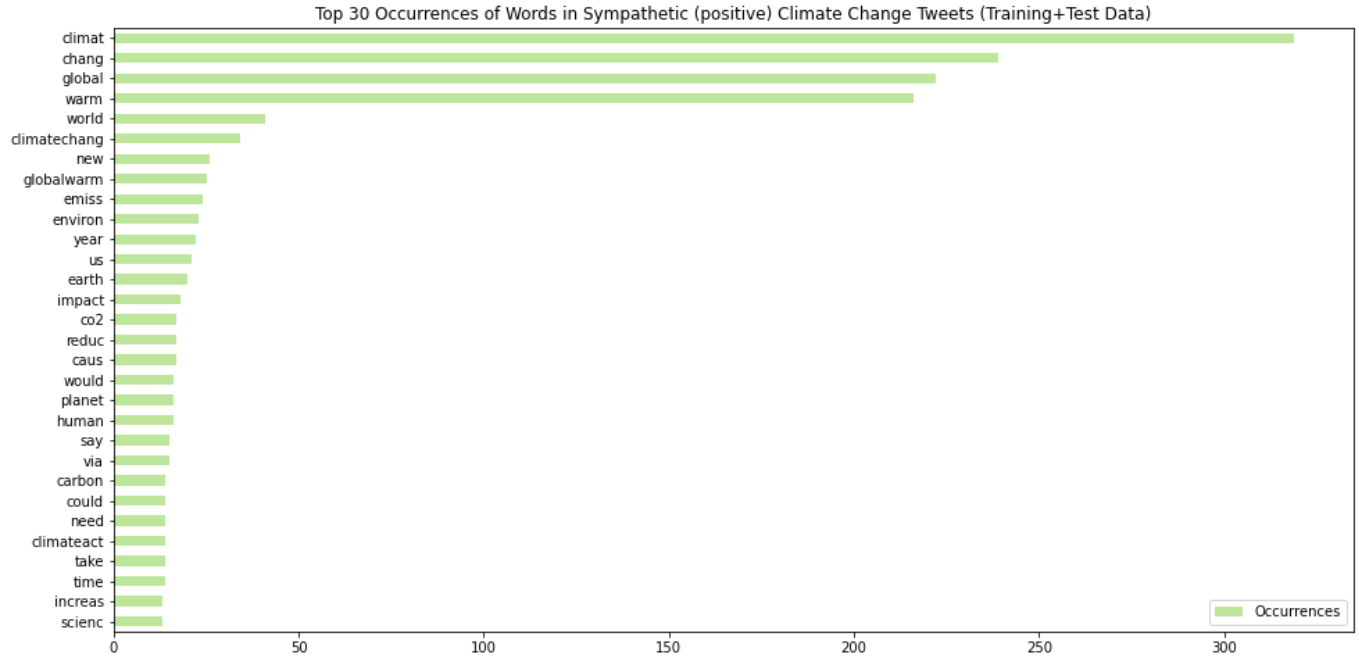

Figure 2. Bar graph of word occurrences in the positive tweets of the dataset used to train the machine.

The words shown seem incomplete because they are determined by the tokenizer to be the root word, e.g. "chang" can mean "changing", "change", "changes", etc.

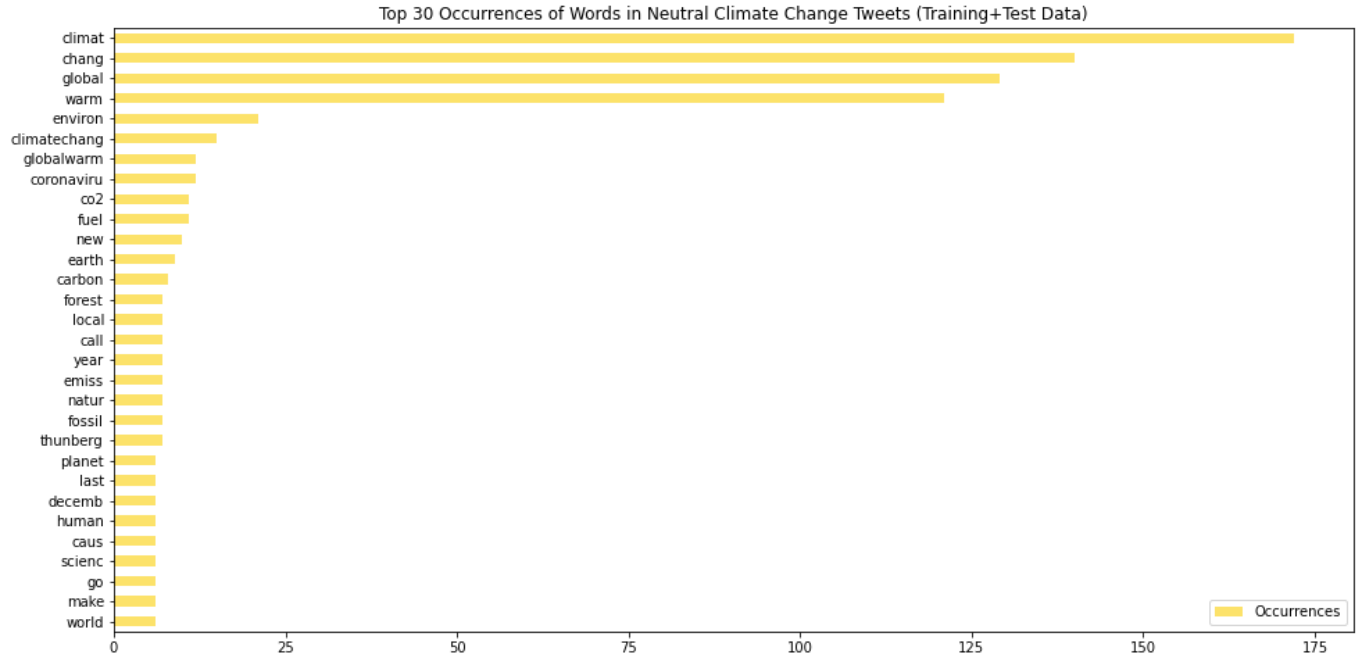

Figure 3. Bar graph of word occurrences in the neutral tweets of the dataset used to train the machine.

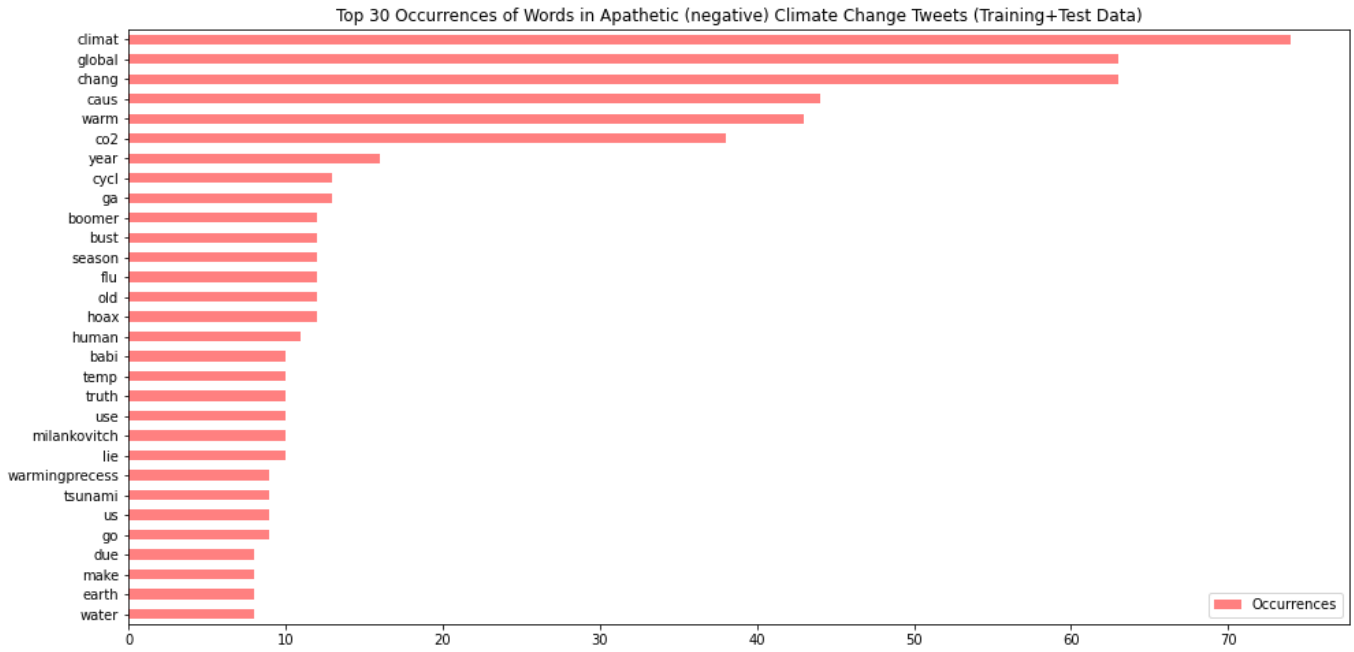

Figure 4. Bar graph of word occurrences in the negative tweets of the dataset used to train the machine.

Accross all sentiments, the most occurring words are "climat", "chang" and "global"; this can be attributed to the reason that the data was scalped from search results made with these keywords.

Hence, we now take a look at the occurrences and wordclouds of the most occurring words, excluding the three most occurring words, "climat", "chang", and "global". The wordclouds are constructed based on the data the bar graphs use, so the reduced bar graphs and wordclouds will be shown consecutively.

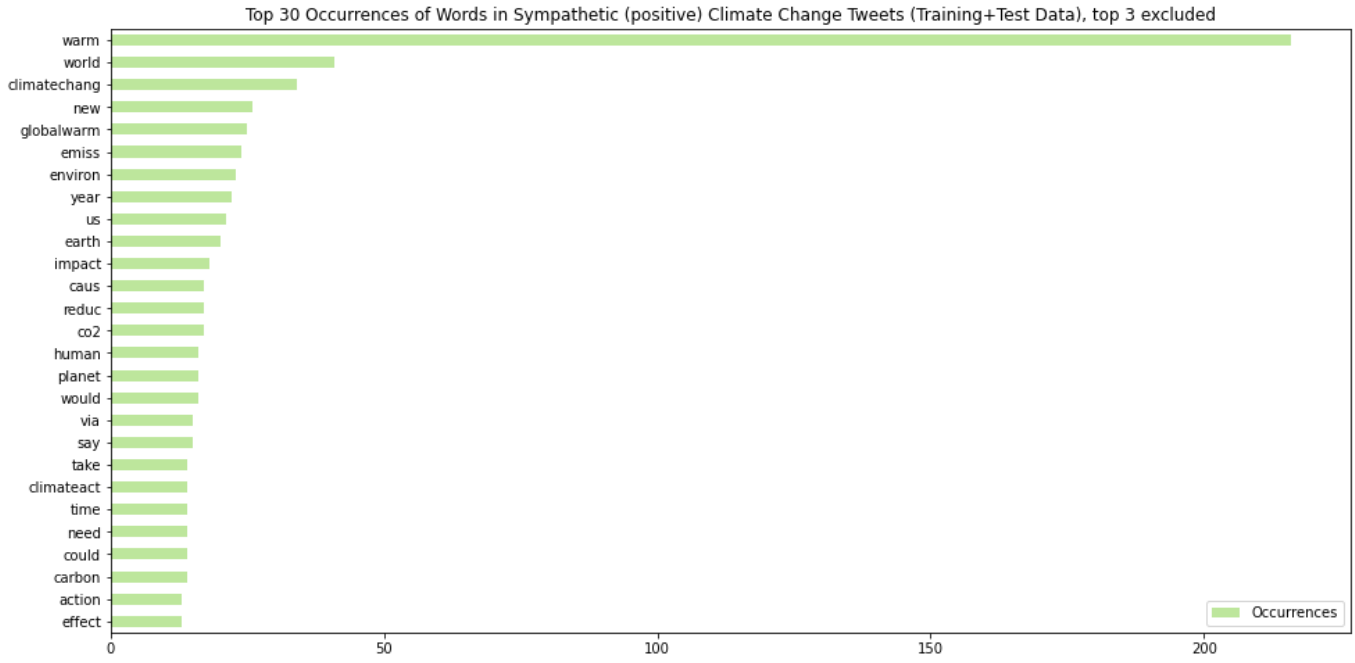

Figure 5. The same bar graph as Figure 2, but excluding the top three words.

Figure 6. Wordcloud of the positive tweets of the training dataset.

The wordcloud of positive tweets shows that words associated with climate change sympathy are that of urgency, science, time, and recognition of the issue as a negative one. Figure 6 thus implies that sympathetic tweets talk about the urgency of solving the issue of climate change.

Notable words that appear in the wordcloud are "reduc[e]" and "emiss[ions]", "caus[e]", "couldfight", and "human". These words imply that climate change sympathetic tweets show optimism and recognition of the cause of climate change, as well as suggest possible solutions for the issue.

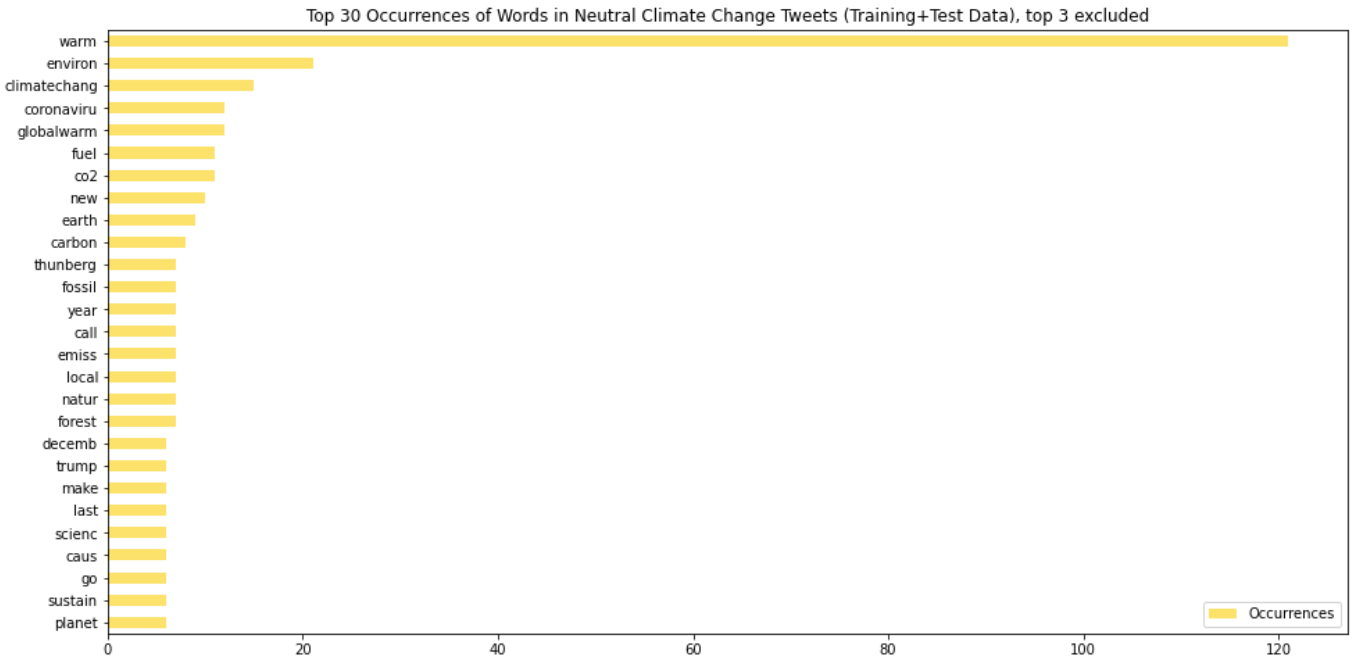

Figure 7. The same bar graph as Figure 3, but excluding the top three words.

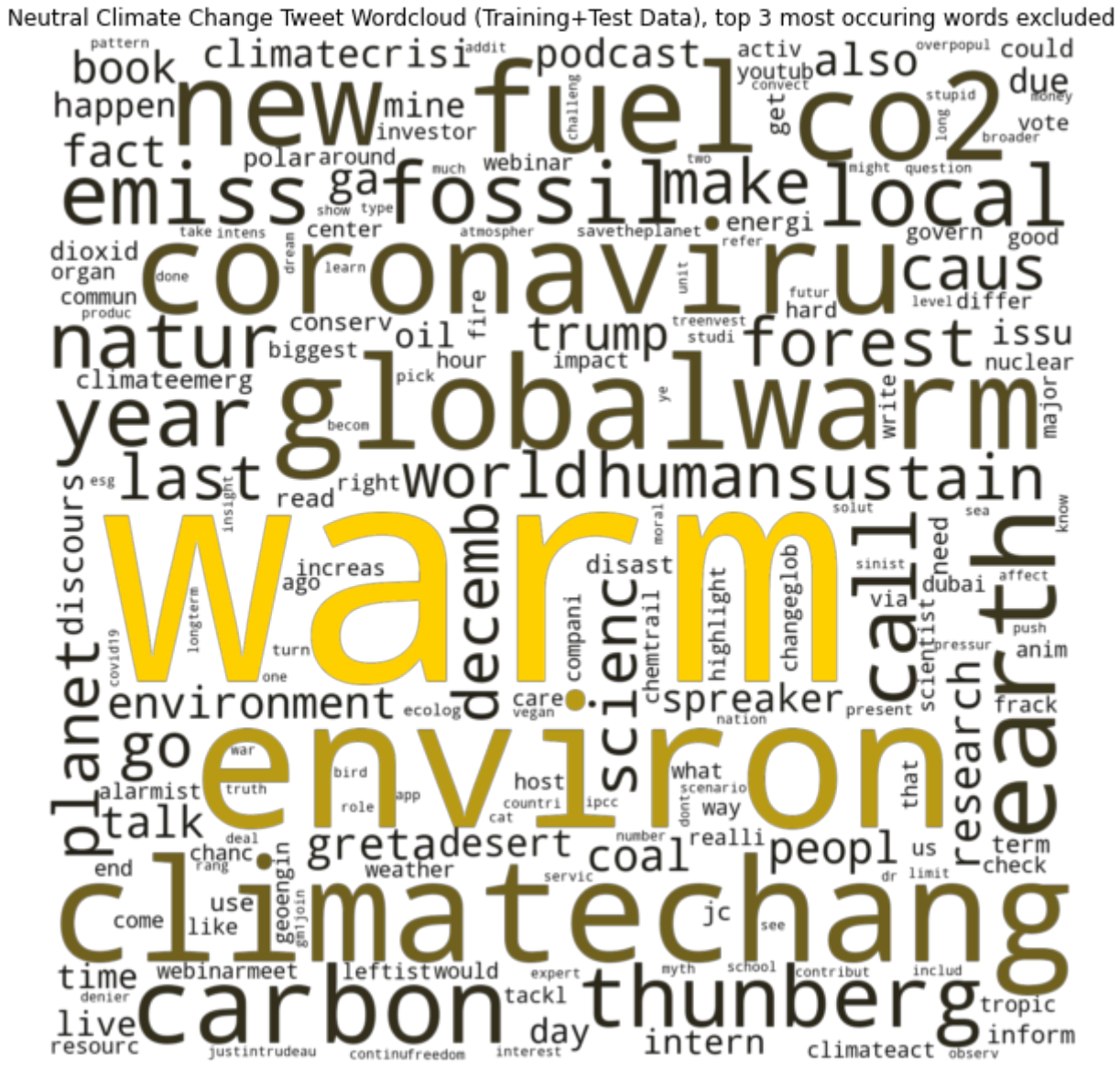

Figure 8. Wordcloud of the neutral tweets of the training dataset.

The wordcloud of neutral tweets shows that words associated with climate change neutrality are associated with concern for other global events and common causes of global warming (particularly fossil fuels). This implies that the neutral climate change sentiment is associated with general awareness of the issue.

Notable words that appear in the wordcloud are "thunberg", "local", "coronavirus" and "new". Greta Thunberg is a Swedish climate change activist that grew in popularity due to her initiative and age. (BBC, 2019) The dataset is composed of tweets from 2020, which was the time when the Coronavirus pandemic struck the world.

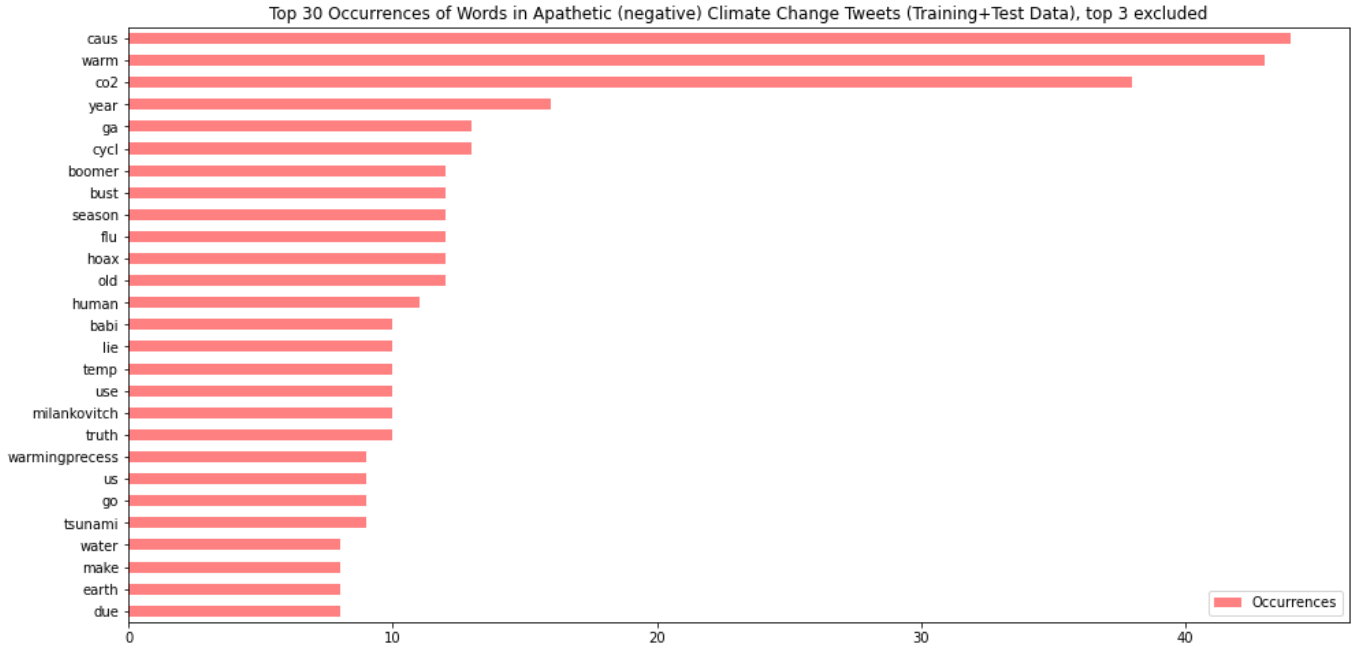

Figure 9. The same bar graph as Figure 4, but excluding the top three words.

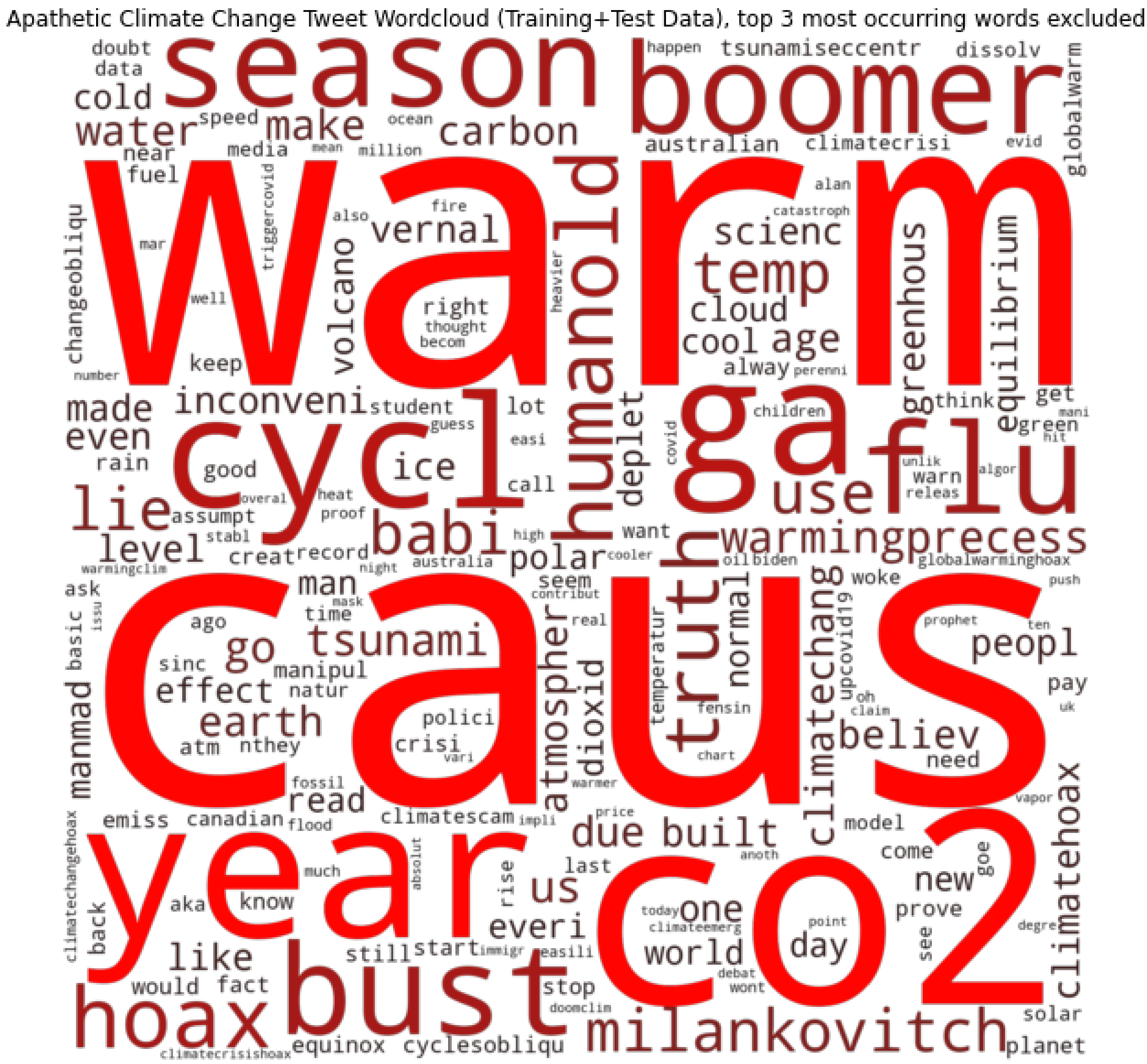

Figure 10. Wordcloud of the negative tweets of the training dataset.

The wordcloud of apathetic tweets shows that words associated with climate change apathy are associated with concerns regarding the truth of climate change and its causes. This implies that tweets apathetic to climate change express belief in the facts that proves climate change is the cause of natural occurrences.

Notable words include "milankovich", "boomer", "hoax", "old", "season", and "cycle". The Milankovich cycle relates the effects of the earth's position, rotate and orientation with the changes in climate. "Boomer" is a term used to refer to the Baby Boomer generation, born during the post-WWII baby boom, and are often prone to misinformation from fake news. (Brashier & Schacter, 2020) This implies that climate change-apathetic tweets express, or are possibly caused by, misinformation.

Conclusion

This study has determined that, in the past half decade, there has been a significant positive change in sentiment on Climate Change.

The training data used for this project was too small to be truly certain of the change in climate change sentiment. It is recommended that, in future replications of this project, the data is larger, and that the data has equal distributions among the positive, neutral, and negative tweets.

References

(2022) Twitter Usage Statistics. *Internet Live Stats.* Archived 2022, May 7, 5:55 PM GMT: https://web.archive.org/web/20220507054131/https://www.internetlivestats.com/twitter-statistics/

Guzman, J. (2020). 2020 Climate Sentiment on Twitter. Kaggle. Retrieved from https://www.kaggle.com/datasets/joseguzman/climate-sentiment-in-twitter?resource=download

Qian, E. (2019, November 13). Twitter Climate Change Sentiment Dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/edqian/twitter-climate-change-sentiment-dataset

Kim, R. (2022, April 5). Twitter Climate Change Analysis. Kaggle. Retrieved from https://www.kaggle.com/code/roellekim/twitter-climate-change-sentiment-analysis

BBC. (2019, April 23). Climate change activist Greta Thunberg: 'Listen to climate scientists' (Video). *BBC News.* Retrieved from https://www.bbc.com/news/av/uk-48018034

Brashier, N. M. & Schacter, D. L. (2020, May 19). Aging in an Era of Fake News . *Current Directions in Psychological Science, 29(3).* doi:10.1177/0963721420915872