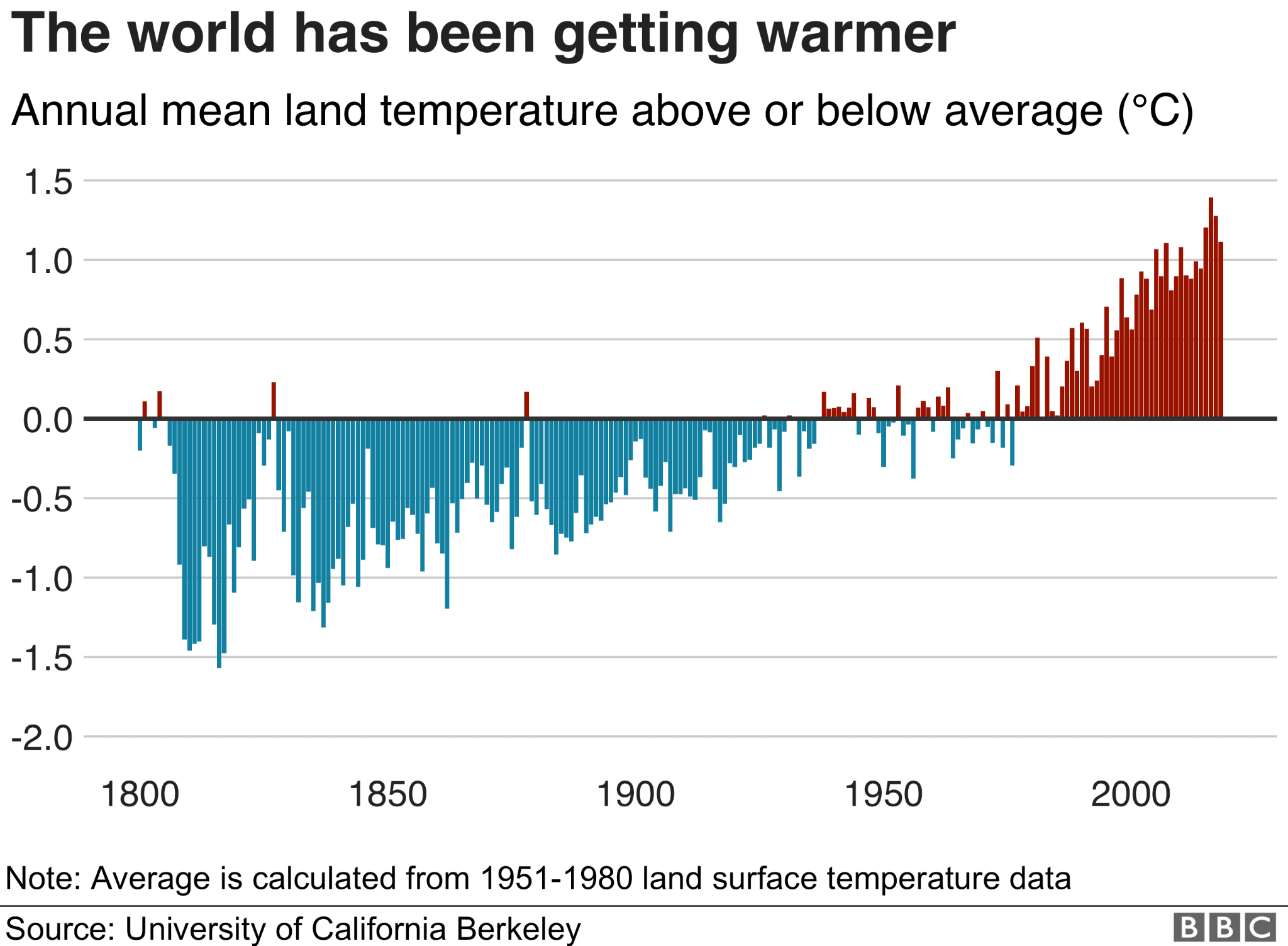
Text Mining Final Project:

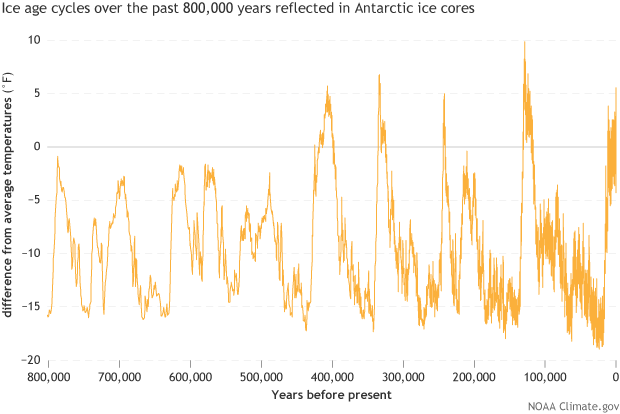
Climate Change

# Introduction

Climate change, also known as global warming, is a growing concern for many and can be a polarizing and politicized subject. Generally, it is the identification of a trend of increasingly warming temperatures, most notably in the past 50 years. Many feel that as this trend continues, life on Earth (plants and animals alike) is sure to face consequences as dire as our very survival.



The cause and concern for this general warming is disputed and discussed across a range of positions. Some data suggest that, in looking at more longitudinal representations like hundreds of thousands of years, this warming is part of a naturally occurring cycle and that the world will recalibrate on its own. Some even point to similar climate events like the Little Ice Age[[1]](#footnote-0) following the Medieval Warm Period. Another school of thought claims this climate fluctuation is likely due to the industrial revolution, population growth, and pollution, while others believe the data is part of a hoax or conspiracy theory.



[Increases and decreases in Antarctic temperature relative to the recent (1,000-year) average during the naturally occurring ice ages of the past 800,000 years, ending with the early twentieth century. NOAA Climate.gov graph by Fiona Martin, based on](https://www.climate.gov/sites/default/files/PaleoTemp_EPICA_large.png) EPICA Dome C ice core [data](http://www.ncdc.noaa.gov/paleo/metadata/noaa-icecore-6080.html) provided by the Paleoclimatology Program at NOAA’s National Centers for Environmental Information.[[2]](#footnote-1)

Potential issues resulting from this warming include a myriad of concerns, including:

* Increase in the severity and frequency of extreme weather events like floods, hurricanes, heatwaves, and wildfires.
* Impacts to food security as droughts or other extreme events cause reduced crop yield; Rise in world hunger and access to fresh water.
* Risk of armed conflict due to lack of resources, and economic impacts as prices increase on food and consumer goods.
* Continued concerns about air quality and health, due to heat waves, pollution, and poor air quality from wildfires.
* Environmental impacts like disruption to the balance of ecosystems, like species extinction from lack of resources or competition from invasive migrating species. Warming ocean temperatures lead to coral bleaching, and reduced oxygen levels for marine life.

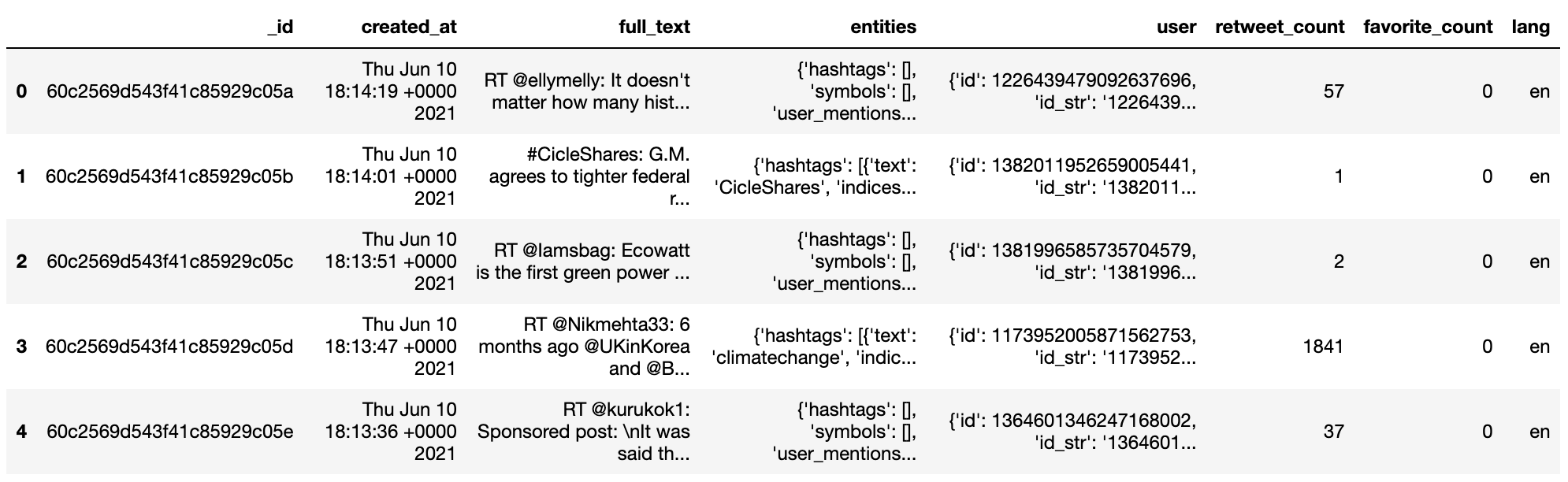
# Analysis

## About the Data

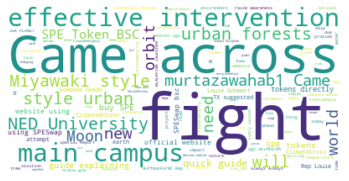
The data is a collection of Tweets, scraped using an API with the Tweepy Python package. Hashtags relating to climate change; global warming; as well as hashtags in opposition to these supposed phenomena were used, including: #climatechange, #globalwarming, #climatechangefraud, and #globalwarminghoax. In total, a collection of 2,066 tweets were captured. For the purposes of text analysis, focus was given to English tweets, which totalled 1,907 of the original collection.

A second task concerning potential solutions to this problem was also explored. Hashtags #climatesolutions and #climatepolicy were collected, totalling 645 tweets. Of these, 572 in English.

Tweets collected via the API were stored in a MongoDB, and output to .csv format for further processing, like removing certain characters and stopwords. The data was then formatted into a pandas dataframe, simplified into eight columns.



18 languages are represented in the data (Twitter supports 34 languages in total). Most Tweets are in English (1907). Other languages include French, Urdu, Hindi, Italian, and Spanish. The language could not be identified for 84 tweets.



After removing stopwords and obvious terms such as those present in the hashtags themselves, a word cloud demonstrates frequently used terms like ‘fight’, and ‘intervention’ while phrases like ‘Came across’ and ‘NED University’ were also common due to popular retweets (RT). The total collection of words for this image comprised 317,392 words.

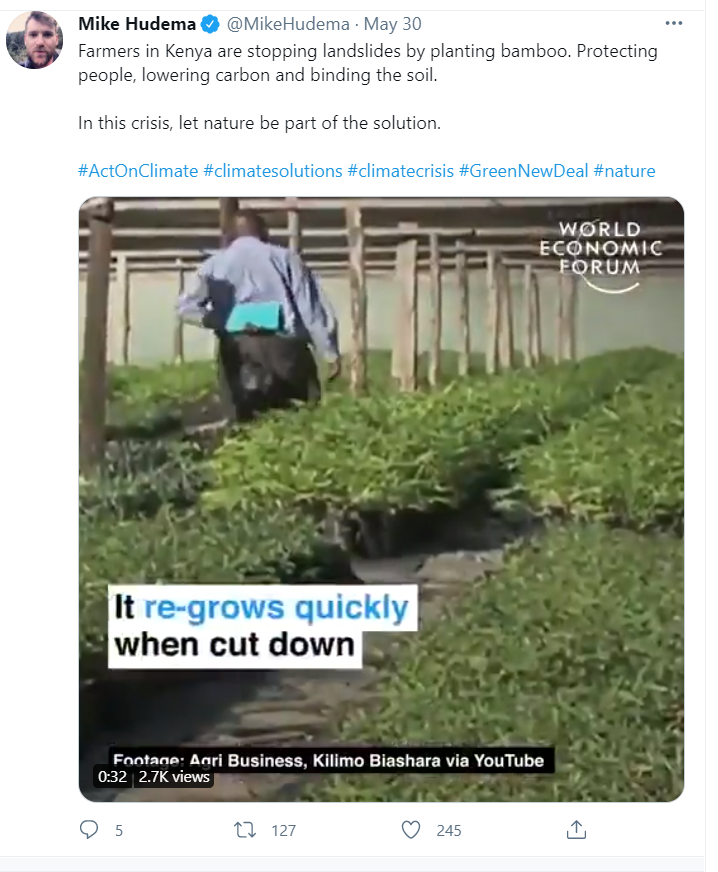
|  |  |
| --- | --- |

Most Favorited, and Most Retweeted tweets, respectively.

The first, discussing a possible solution of urban forests, a method of intervention as promoted by NED University; and the Korean Pop group BLACKPINK, using their platform to raise global awareness on the issue of #climatechange.

## 

## Additionally, after removing english stop words, single letter words, and other words that would be meaningless to the analysis, the climate solutions dataset word cloud highlighted terms like “Futures Act”, “ensure”, “bill”, “cosponsor” and “Clean Futures”. These terms appear to be focused around a call for congressional act as a solution.



## The two most favorite and the most retweeted tweets are pictured above. The most retweeted tweet coming from Project Drawdown is advertising a free online course centering on game-changing climate action, and discusses the idea of need-to-know science in Climate Solutions. It was retweeted 169 times. The most favorite tweet came from Twitter user Mike Hudema, highlighting a video from the World Economic Forum where Kenyan farmers are taking their own action to stop landslides. This user took this opportunity to equate these actions with these farmers doing their part to be part of the solution for climate change, whether that was their intention or not.

## 

A distribution of the top 20 most frequently occurring words also highlights words centered around congressional action.

## Models

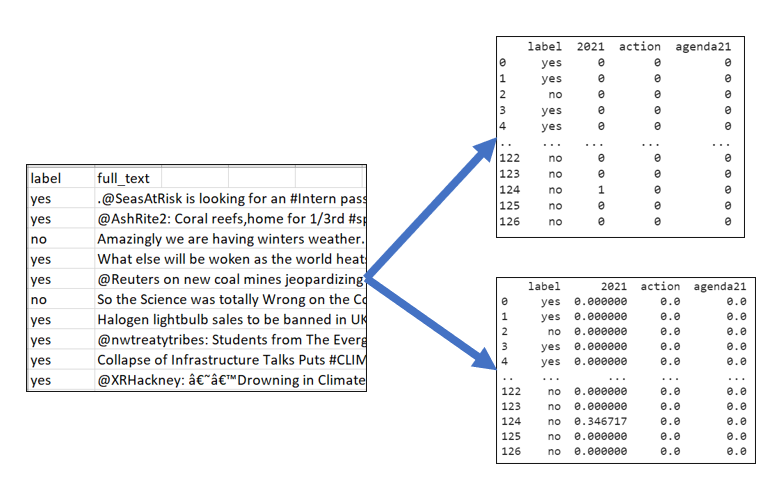
### Sentiment Analysis

One method of analyzing Twitter data is by looking at the sentiment of different tweets. By analyzing the sentiment of different tweets and understanding if they are positive or negative, the general feeling towards climate change can be gauged. Rather than hand label this data, sentiment can be easily derived using a tool called Vader. Vader is a part of the nltk Python package and with a text input will return a positive, neutral, and negative sentiment score from zero to one - and most importantly a compound score between negative one and one, with one being the most positive and negative 1 being the most negative. By using this compound score, overall sentiment can be determined using the logic that a positive score represents positive sentiment and a negative score represents negative sentiment.

In order to retrieve results for this dataset, the data does not need to be tokenized or otherwise prepared for the Vader algorithm to work. Instead, the column in the dataset that contains the text can be passed in its original form to be processed by the algorithm. This makes the algorithm a very simple way to extract sentiment from text data, since it is unsupervised and requires very little data processing. By analyzing these scores the general sentiment towards climate change based on these tweets will be distinguishable.

### Climate Change Believers & Skeptics

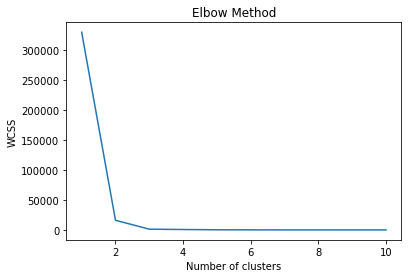
An assessment of the differences and similarities in how climate change believers and skeptics tweet about climate change was done using classification. To further prepare the data for this analysis, a subset of the data containing 127 tweets was created for hand labeling. The labels applied were either a “yes” (believer) or “no” (skeptic). Because this dataset is heavily weighted towards believers in climate change, it took multiple subsets to achieve enough climate change skeptic tweets to move forward.

Once labeled, the data was loaded into Python for vectorization. For comparison purposes, two vectorized data frames were created: one using CountVectorizer for word frequency vectorization and the other using TfidfVectorizer for TF-IDF weighting vectorization (seen in the figure below). Other parameters applied were the removal of stopwords, a minimum document frequency of 2, and a minimum word features of 1,000. Certain features with digits or special characters were also scrubbed from the data.

For the classification task itself, two models were built using the multinomial Naive Bayes algorithm applied to the two vectorized data frames. Naive Bayes was chosen due to the relative ease of interpretation of the model.

### Climate Change Solutions Clustering

After hypothesizing there could be clusters of conversations within the climate solutions dataset, a cluster analysis was run on the data using Python’s sklearn library. The data was vectorized by Tfidf and Count Vectorizer and tested on a KMeans algorithm. The algorithm was tuned with *k-means++* as the method of initialization and 500 max iterations to fit the data. The silhouette “elbow” method determined there were three clusters of vocabularies within the data.

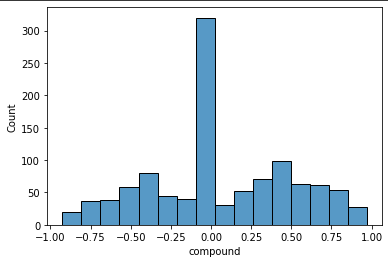


# 

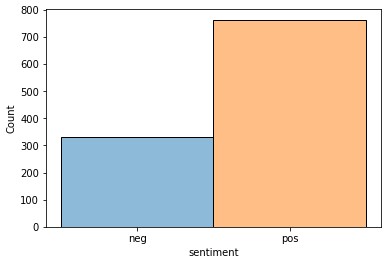
# Results

## Sentiment Analysis

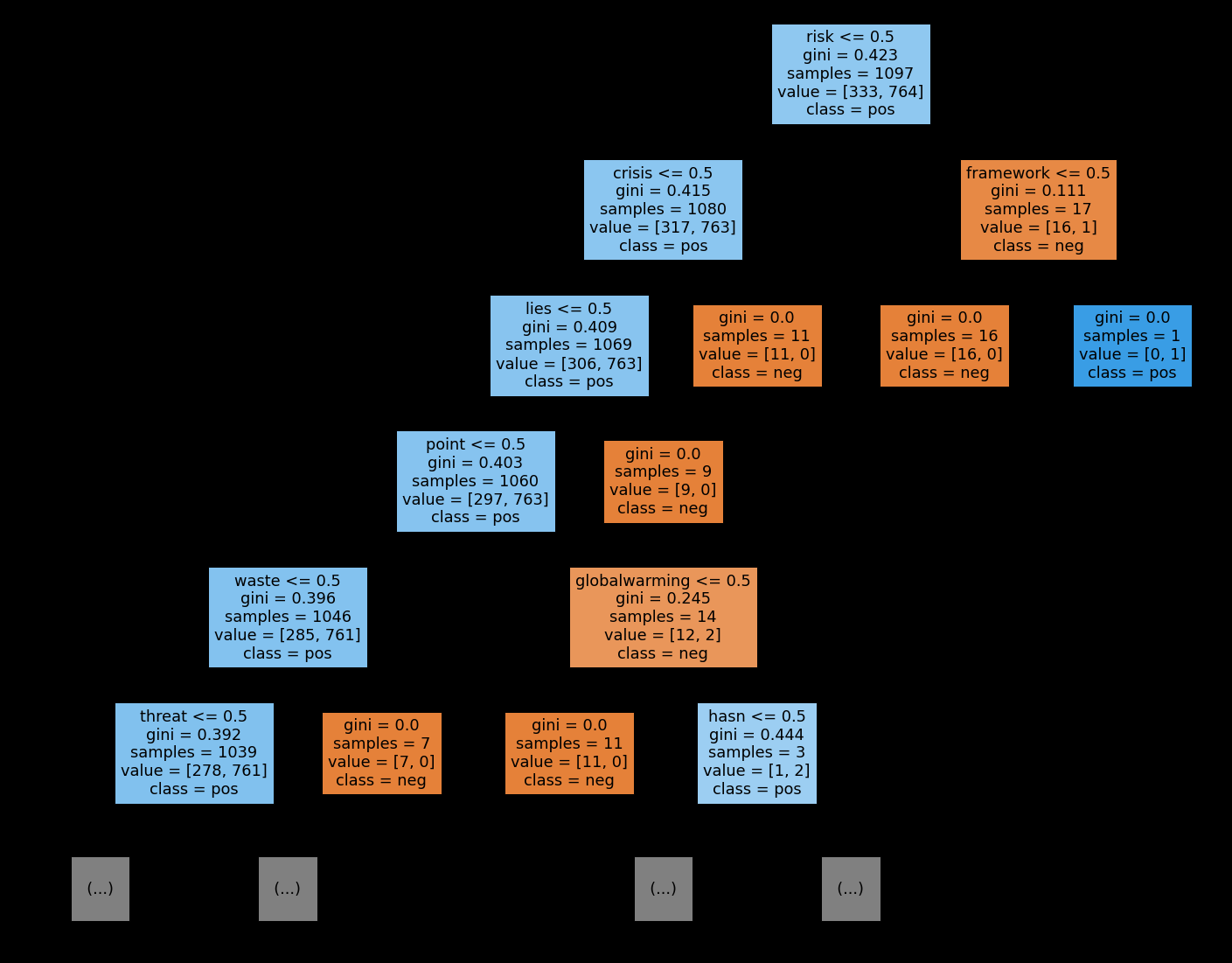
The results of the Vader analysis showed generally positive sentiment regarding climate change, with a large proportion of the data having near-zero sentiment - likely being more informative tweets which do not take a stance in either direction. The distribution of scores was as follows:



These results showed over 300 of the tweets with a near-zero sentiment score. These tweets were mainly classified as positive, since there was zero or very slightly positive. Due to this, there were far more tweets with a positive sentiment than a negative sentiment, which can be seen below:



This was an expected result because the tweets collected contain more tweets with hashtags that would be considered pro climate change or used by people who believe climate change to exist. Lastly, by using a decision tree and count vectorizer to attempt to predict the sentiment, as determined by Vader, the resulting tree will show what unigrams were most important in determining sentiment to help give some insight into how Vader compiled the sentiment scores. The top six levels of the tree showed the following results:

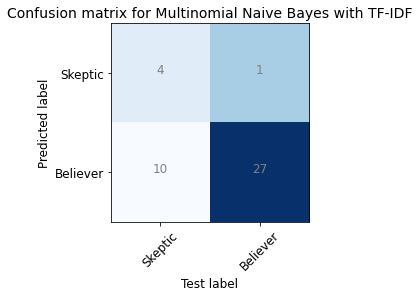


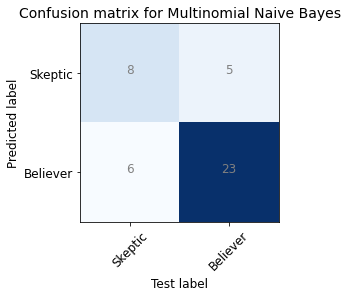
As seen by the tree, the word “risk” was the most important word for determining sentiment; and people who used the word “risk” at least once are likely to have a negative sentiment. Then, looking at the rest of the tree, some of the most important unigrams were ones such as “crisis”, “lies”, “threat”, and “globalwarming”. In this case, “globalwarming” occurs as a unigram since it was part of a hashtag in the data and would not be split by the vectorizer. By using Vader, it was clear that the majority of tweets were neutral to positive, which aligned with the tweets that were collected. Then, the unigrams which were best for separating positive and negative tweets, based on Vader's results, and the most important were words that people would often use when debating climate change or global warming.

## 

## Climate Change Believers & Skeptics

|  |  |
| --- | --- |

The Naive Bayes models were both trained on 66% of the dataset and tested on 33% using a hold-out test. Both models scored a 74% accuracy (31 out of 42 correct). The errors were different with each model (see figure below). The TF-IDF model scored almost all skeptics as believers. Compared to the TF-IDF model, the word frequency model scored more believers as skeptics. This low accuracy could potentially be due to the skew in the data mentioned in the previous section, as well as the small dataset.

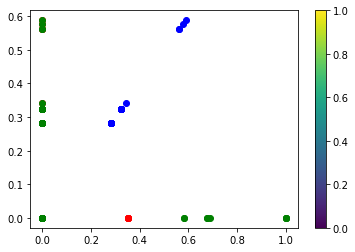


The most indicative words of a climate change believer and a climate change skeptic tweet gives insight into what the model learned from the data (see two figures below). There were a few observations to call out. In this dataset, there seems to be a difference in terminology between the two groups. “Climate” and “change” appeared as indicative of a climate change believer, whereas “warming” was indicative of a climate change skeptic. Further, other indicative words of a climate change believer's tweets were active verbs, like “fight”/”fighting”, “join”, and “action”. “COP26” appeared in the words indicative of a climate change believer tweet. This is an international climate change conference occurring in a few months. “Drought” was another word in the list, which could be referencing a current climate emergency in California.

Other than “warming,” there were a few other interesting observations from the list of words most indicative of a climate change skeptic tweet. “Greta” and “Thunberg” appeared in the list. Greta Thunberg is a young climate activist who has gained media attention in the last few years. “Data” and “science” also appear in the skeptics list, along with “Covid-19,” “virus,” and “pandemic.” Another interesting term in the words indicative of a climate change tweet was “Agenda21,” which refers to a UN resolution from 23 years ago. There also appears to be a conspiracy theory surrounding this UN resolution, although it is unclear in what context the term is being used here. Finally, “cold” and “weather” appear in the indicative words list for climate change skeptics tweets.

## Climate Solutions

The results of the cluster analysis showed a cluster with highest frequency of words like “bill”, “clean”, “cosponsor”, “ensure”, “floor”, “futures act”, “please” and “vote”. This cluster was centered around congressional action. Whereas another cluster with highest frequency of words like “carbon”, “climatesolutions”, and “drawdownga”. These were much more about promoting change through individual action. “Dawdownga”, or Drawdown Georgia, is a movement gaining traction in Georgia to accelerate progress toward net zero greenhouse gas emissions, with a reduction of at least one-third by 2030.

The cluster graph below highlights the 3 clusters discovered within the data. The green cluster (cluster 1) had the strongest weighted vocabulary. These were the words requesting congressional action as a solution to climate change. The red cluster (cluster 2) had very low rated vocabularies and few words belonged to this cluster. Lastly, the blue cluster (cluster 3) separated from the rest of the data indicating a clear delineation of words from the other clusters with vocabularies centered around individual action as a solution using words like drawdownga and carbon.

# Conclusions

Public sentiment on Twitter related to climate change and global warming seems to be largely neutral to positive. Those tweeting negatively about climate change seem to be voicing concerns about risk, crises, and threats. There also seem to be many people tweeting informatively about climate change, rather than voicing an opinion one way or another.

There are differences in the ways climate change believers and skeptics tweet about climate change. It appears climate change believers are more likely to use the term “climate change” compared to climate change skeptics. Climate change skeptics are more likely to use the word “warming.” Climate change believers are more likely to use active verbs and calls to action, whereas climate change skeptics are more likely to bring up science and data. Climate change skeptics are also more likely to mention climate change and the COVID-19 pandemic in the same tweet.

Both climate change believers and skeptics are likely to talk about weather events, but climate change believers are more likely to talk about droughts (or potentially other current or recent extreme weather events), whereas climate change skeptics are more likely to use terms like “cold” and “weather” (and potentially the two terms together). This gives more insight into the different arguments, feelings, and beliefs climate change believers and climate change skeptics hold about climate change.

The words used in tweets about climate solutions call for congressional and personal action. Climate change solutions tweets that focus on policy change use words to encourage politicians to enact bills and laws for clean futures. Climate change solutions tweets that focus on personal action use words like “carbon” and “drawdownga”. These words demonstrate the solutions that Twitter users believe will be effective at mitigating climate change.

1. Mann, M. E.; Zhang, Z.; Rutherford, S.; et al. (2009). ["Global Signatures and Dynamical Origins of the Little Ice Age and Medieval Climate Anomaly"](http://www.geo.umass.edu/climate/papers2/Mann2009.pdf) (PDF). Science. 326 (5957): 1256–60. [↑](#footnote-ref-0)
2. [climate.gov/news-features/climate-qa/whats-difference-between-global-warming-and-climate-change](https://www.climate.gov/news-features/climate-qa/whats-difference-between-global-warming-and-climate-change) [↑](#footnote-ref-1)