ReAcclimate: The New Climate Change Lexicon

Rui Aguiar [raguiar2@stanford.edu]
Anthony Carrington [acarring@stanford.edu]
Harold Wang [haroldw@stanford.edu]

Overview

Problem Definition: In the 1980s, the Republican party hired a team of top linguists to define a lexicon of terms that would drive engagement from both parties and swing votes towards their viewpoints. We wanted to answer the question: Can we use twitter data to define a lexicon of words that will drive cross-platform discussion about climate change? We leveraged K-Means clustering and RNN's with LSTM cells to answer this question

Example output:
As an example, our model may suggest the following changes in language

"fight" -> "crusade" "law" -> "justice" "earth" -> "home"
"case" -> "lawsuit" "flood" -> "disaster" "ocean" -> "sea"

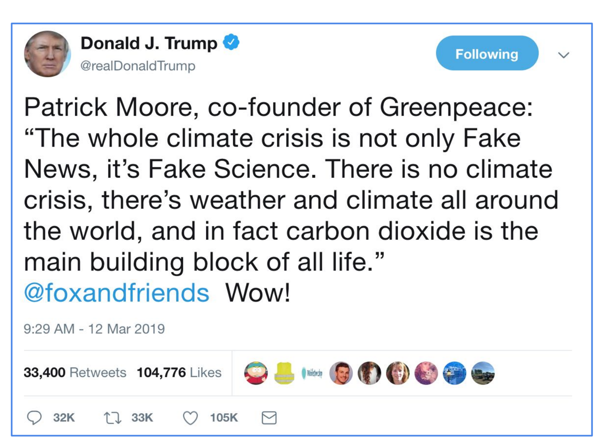
Data Source

Twitter: We used Tweepy, a Python Twitter API wrapper, to collect over 2M climate change tweets created between September 21, 2017 and May 17, 2019 that were found using the following keywords and hashtags: climate change, global warming, climate hoax, #climatechange, #climatechangeisreal, #actonclimate, #globalwarming, #climatechangehoax, #climatedeniers, #climatechangeisfalse, #globalwarminghoax, #climatechangenotreal

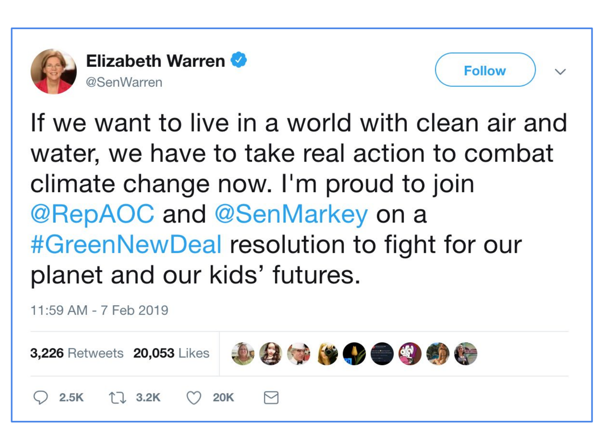
Relevant Features:
- Tweet text
- Engagement: # of favorites + 3 * # of retweets
- User screen name
- User follower count
- Text sentiment polarity

Preprocessing:
- Converted to lowercase
- Spaces added between concatenated words (as in hashtags)
- Punctuation stripped (@, # symbol, etc.)
- URLs removed
- Emojis and emoticons removed
- Contractions expanded to two words

Examples:



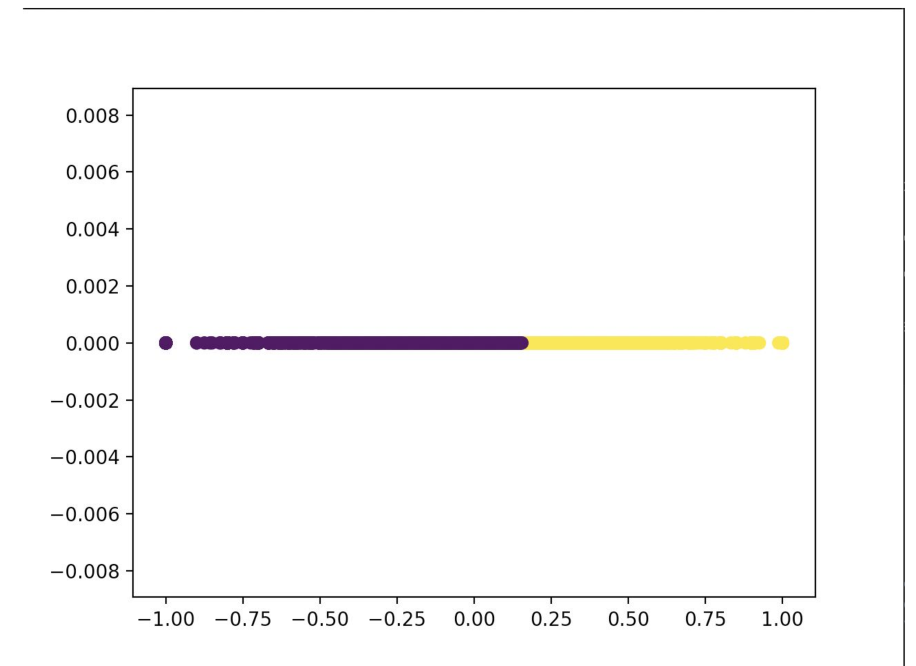
Text: patrick moore cofounder of greenpeace the whole climate crisis is not only fake news it is fake science there is no climate crisis there has weather and climate all around the world and in fact carbon dioxide is the main building block of all life wow
Sentiment polarity: -0.089
Follower Count: 60,883,582
Engagement: 204,976



Text: if we want to live in a world with clean air and water we have to take real action to combat climate change now i am proud to join and on a green new deal resolution to fight for our planet and our kids futures
Sentiment polarity: 0.232
Follower Count: 5,051,195
Engagement: 29,731

Unsupervised Learning

K-Means Clustering: To determine who our groups were to predict cross-group engagement, we decided to use unsupervised learning. We experimented with several features including subjectivity and polarity, before deciding to group on average sentiment of a user towards climate change.



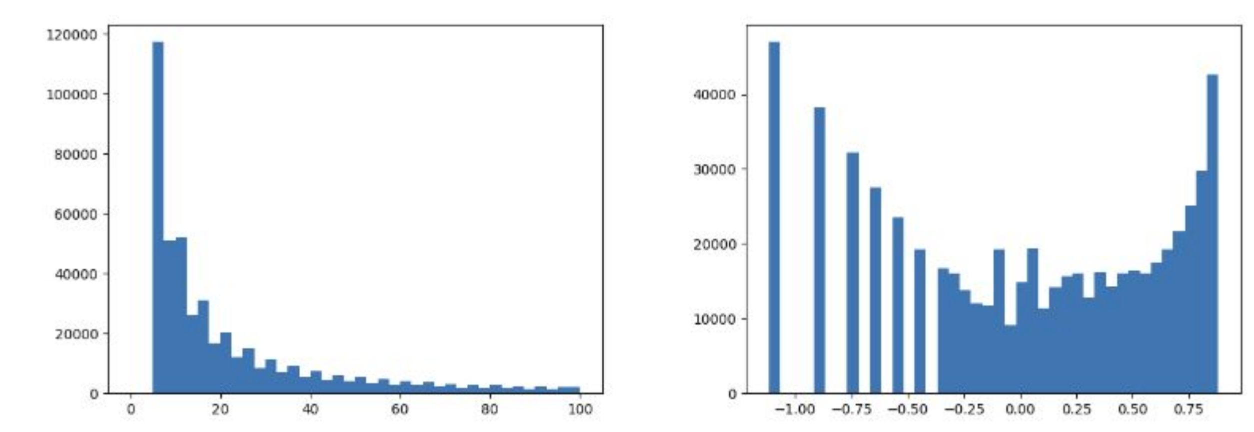
One output from our k-means clustering with k=2

Model Output: For our K-Means algorithm, we eventually settled on k=2 for our grouping, and to cluster on average sentiment. The reason behind this was that we were able to get relatively sizable groups, and a users tweets tended to have similar sentiment.

Engagement Prediction

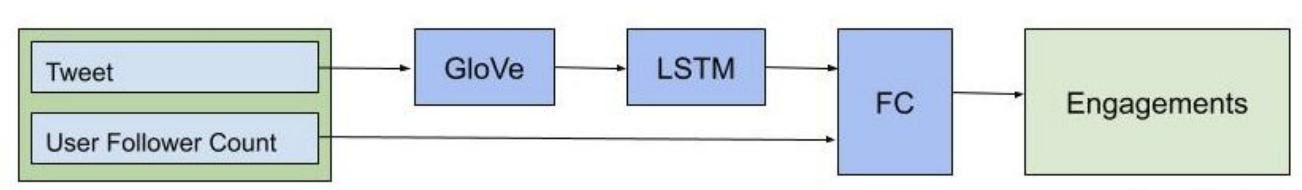
We calculated the user engagement as a function of the number of favorites and the number of retweets. We applied a Box-Cox transformation to obtain a uniform engagement distribution to improve our system performance.

$$E(N_{reply}, N_{retweet}) = \frac{-1}{0.6}((N_{reply} + 3 * N_{retweet})^{-0.6} - 1)$$



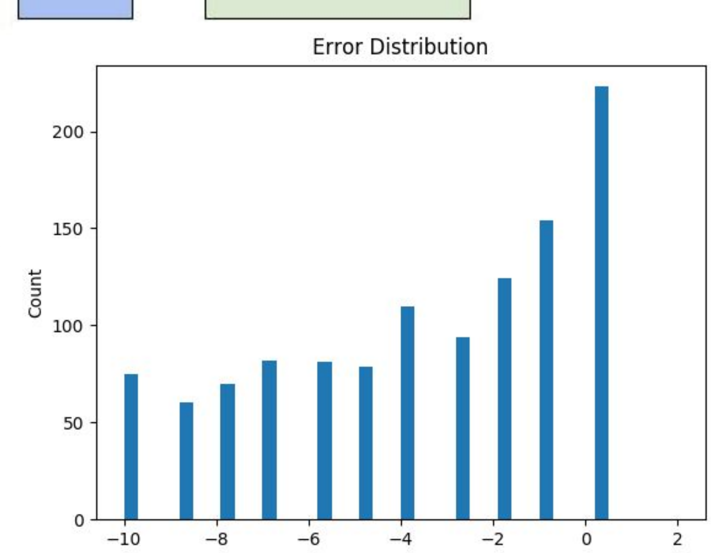
(a) Original Engagement Score Distribution (b) Transformed Engagement Score Distribution

Since users with opposing climate change views may have different responses to a particular tweet, we divided our dataset into classes by user group and trained separate models to predict their engagement score.



Each model has a two layer LSTM network followed by 4 fully-connected layers. The inputs are the embedded tweet text using GloVe[1] word representation and the user's follower count. We used MSE as the loss function.

The test set error distribution is shown in the figure below.



Synonyms/Results

We found that the following words have the highest impact on user engagement score.

Engagement Delta	Original Word	Suggested Word
1792	threatens	endanger
1536	approval	blessing
1280	immediate	contiguous
1280	contaminated	pollute
1024	costing	cost
1024	spare	bare
1024	greening	rejuvenation
896	options	choice

Analysis and Future Work

Word Recommendations
- Words that are positive, empowering, and hopeful in nature, such as: "blessing," "rejuvenation," and "choice"
- Words that are tied to uniquely negative, human-caused phenomena such as "endanger" and "pollute"

Data Improvements
- Beneficial to gather more tweet reply data
  - If a climate change "hoax" proponent responded positively to a tweet addressing the climate crisis, this would likely indicate that the parent tweet used really great language. Further investigation here is needed
  - We were limited in collecting this data as it is not well supported by the Twitter API

Input Features
- Sentiment is not the only way to cluster users, and may not be a reliable indicator of pro vs. anti-climate change disposition.
- Prediction might benefit from more user features, such as demographic data (location, etc.) and political views

Model Overfitting
- Better address how engagement score distribution is heavily skewed towards 0
  - Currently corrected by transforming it into a uniform distribution
  - More data may solve this without a transformation needed
- Address differences in distribution and occurrence frequency of vocabularies between training, validation, and test sets

References

[1] Pennington, Jeffrey. GloVe: Global Vectors for Word Representation, Stanford NLP, Aug. 2014, nlp.stanford.edu/projects/glove/.