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: #globalwarming, #climatechangehoax, #climatedeniers, #climatechangeisfalse, warming, climate hoax, were found using the following keywords and hashtags:

- #globalwarming
- #climatechangehoax
- #climatedeniers
- #climatechangeisfalse

Preprocessing: We used real great language. Further investigation here is needed

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

Relevant Features:

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

Examples

- Text: If we want to live in a world with clean air and water, we need to take real action to combat climate change now. I am proud to join and on a new climate deal resolution to fight for our planet and our kids future.

  Sentiment polarity: 0.232
  Follower count: 5,951,195
  Engagement: 29,731

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

<table>
<thead>
<tr>
<th>Engagement Delta</th>
<th>Original Word</th>
<th>Suggested Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>1792</td>
<td>threats</td>
<td>endanger</td>
</tr>
<tr>
<td>1536</td>
<td>approval</td>
<td>blessing</td>
</tr>
<tr>
<td>1280</td>
<td>immediate</td>
<td>contiguous</td>
</tr>
<tr>
<td>1280</td>
<td>contaminated</td>
<td>pollutate</td>
</tr>
<tr>
<td>1024</td>
<td>costing</td>
<td>core</td>
</tr>
<tr>
<td>1024</td>
<td>spare</td>
<td>bare</td>
</tr>
<tr>
<td>1024</td>
<td>green</td>
<td>rejuvenation</td>
</tr>
<tr>
<td>896</td>
<td>options</td>
<td>choice</td>
</tr>
</tbody>
</table>

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.

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.

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