Climate Conversation Topic Modelling

Report

Karo Castro-Wunsc
2024-3-21

Problem Statement

The Federal Government is drafting a new climate policy and needs to make sure they are engaging with the issues and aspects of climate that are most relevant to the citizens and voter base. They are employing several methods to get an understanding of which Topics are important and which have the most presence in public conversation. One method that is being used is to extract topics from public twitter data from all tweets where people discuss 'climate change' or 'global warming'.

The central Research Question here is:
What Topics can be extracted from climate tweets and how can we achieve the best relevance / coherence for those topics?

Data Wrangling

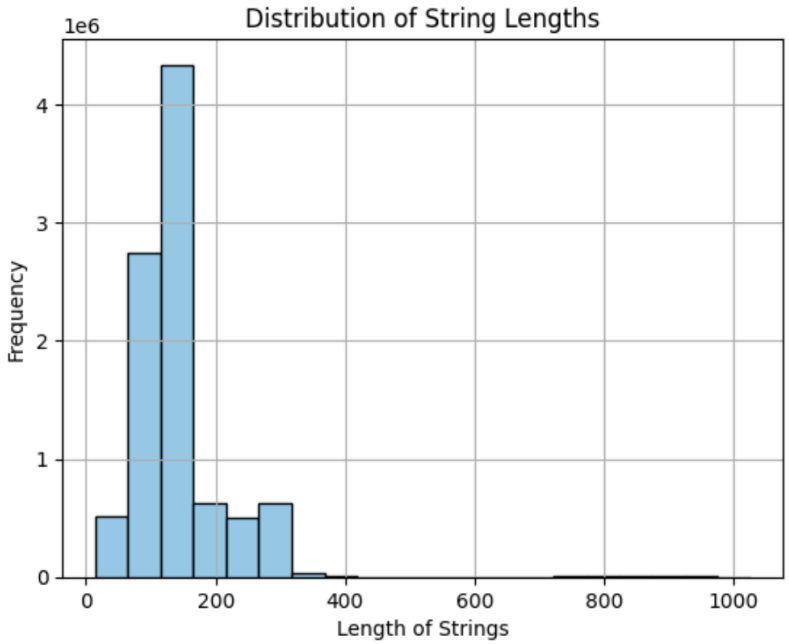
Data Wrangling Notebook

Our primary data source is the twitter firehose data stream archive on the internet archive, we're working with a subset of that data that includes only tweets that contain 'Climate Change' or 'Global Warming'. The data set is primarily 2 columns: 'id' and 'text'. The 'id' in this case is a direct pointer to the tweets canonical id as it is referenced/stored by twitter. The 'text' is a string that contains all hashtags, but no comments. There is no user data. There are roughly 15 million tweets, we will be using a small fraction of these for analysis. The raw 'text' data contains many non-english words. There are references to usernames in the format @{str}, there are emails, urls, emojis and many short-form words and misspellings. It's a chaotic text-space! Tweets have a mean length of 140, with

Steps Taken in Data Cleaning:

1. Drop duplicates. 40% of the dataset is duplicates
2. Investigate tweet lengths
3. Investigate the temporal distribution of the tweets
4. Clean tweet string following standard NLP Conventions

    a. Lowercase
    b. Remove html
    c. Remove Urls
    d. Remove Punctuation
    e. Remove users
    f. Remove hashtags
    g. Remove emails
    h. Remove low info chars (ex. Standalone 'e')
5. We then tokenize the strings
6. We preprocess the tokens
    a. Remove standard english stopwords (ie common, low information words like 'that')
    b. We remove stopwords unique to this dataset ie 'rt', 'amp'
    c. Mild spelling correction for common mistakes
    d. Translate compound tokens ie 'climatechange'

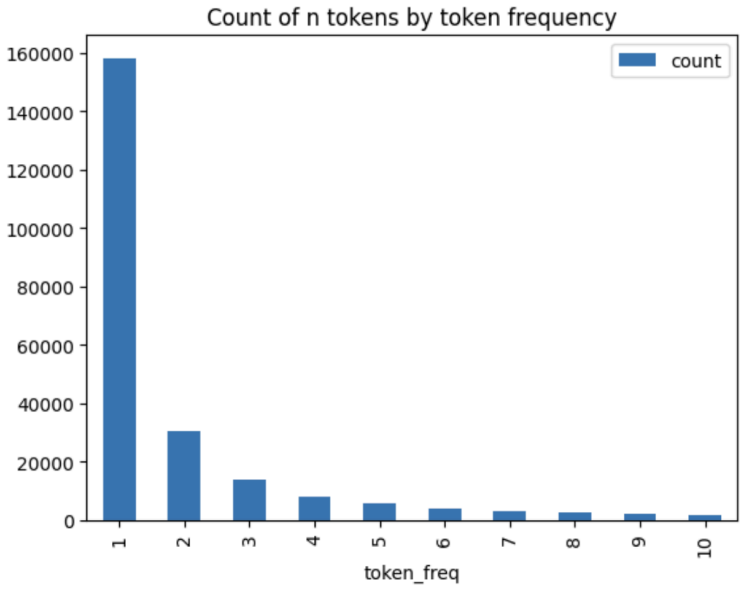


EDA

EDA Notebook

The EDA we perform is with the intention of having the tweet tokens as regularised as possible for vectorization. The tokens have a very broad lexicon and still include many non-english words.
We have 250k unique tokens over the 1m subset of the dataset that we're working with. The vast majority of tokens have a low frequency throughout the corpus. Our goal with token preprocessing is to remove low-info tokens and to normalise the meaningful tokens so that we

don't have disjoint tokens with the same meaning i.e. 'scientist' and 'scientists'. We explore token vectors and tweet vectors which we found to be useful. We also explored TF-IDF on a per-tweet basis but found it to hold little useful structure due to the very low token count per document along with the relative rarity of most tokens.



Token Preprocessing

We took the following steps to pre-process the tokens:

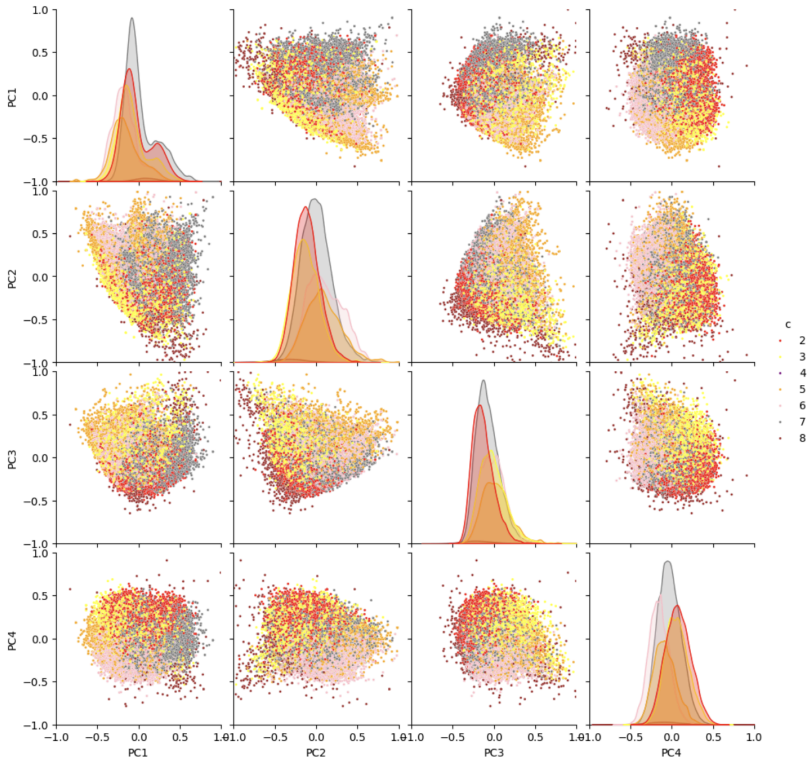
1. We filter out all tokens that occur <= 5 times. This leaves us with a vocab of ~28k tokens.
2. We Lemmatize the tokens to their root forms (4% token reduction)
3. We filter out tokens not in the Word2Vec embedding model. This acts as a 'real world' surrogate for an english dictionary

Token and Tweet Vector Embedding

We explore vector embeddings for the tokens which would give us the ability to represent tokens and tweet strings as a 300 dim vector which is a huge improvement over a 20k long binary word embedding. We use a pre-trained Word2Vec model trained on a very large news headings dataset which we would expect to cover the majority of the important tokens of interest for us. We did not explore Embedding model fine-tuning. We used the vectors for the the tokens of each tweet to generate a Tweet Vector by simply taking the mean of all tokens in the tweet. This is a rudimentary but reasonably supported approach. Conceptually, we would expect this approach to give us a representative 'centre' for the cloud of the tokens of a tweet, but would not necessarily retain specific token info. We tested tweet vec similarity to neighbouring tokens in

the embedding space and found that some vectors did retain specific token associations, which is helpful.

We explored the structure of the token vector embedding space using naive top-word categories and PCA and found that there was significant visual structure retention. This is a good indication as to the usefulness of the tweet vectors.



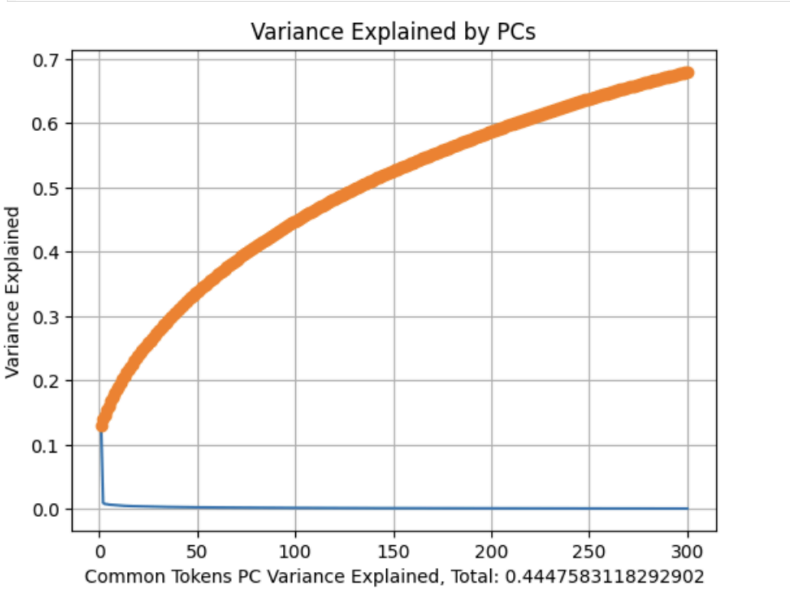
Data Pre-processing and Training

[Process and Train Notebook]

In this phase of our analysis, we: 1. Load and prep token data, 2. Conduct Feature Engineering, 3. Prep Data for Model Ingestion.

We found in EDA that the top tokens covered a large percent of the dataset, we also found that naive 'most significant token in tweet categories' was relatively effective in representing distinct categories in PCA. Based on this intuition, we explore binary vector representations for each tweet that encode the presence of the top N most common tokens in the corpus. We explore using the N=1000 top tokens as this would give 100 tokens per category naively assuming even distribution. We found 1000 dim vectors to till be quite large to work with, so we explored using

PCA to capture as much of the variance of these vectors as possible. We found PCA to have a very 'slow' curve to capture variance per PC included, each PC.



This suggests that using PCs for the features might not be useful for capturing the vector information. We explored SVD as another approach to dimension reduction but with similar results to PCA.
Based on this and visual cluster inspection from PCA we decided this one-hot feature was not a significant improvement over the meaned tweet vectors.

Topic Modelling Model Evaluation

Modelling Notebook

As our ultimate goal in this analysis is to model topics for climate conversation, here we explore the capacity for the explored models to extract meaningful topics. We use several industry standard metric to evaluate the quality of the models we explore. We did a small bit of further feature engineering by incorporating sentence embeddings from a pre-trained Sentence Transformer 'Distilbert'. This gives us an embedding vector similar in meaning to the mean Word2Vec embedding with dims=500. It's possible the sentence vector would retain more information about the sequence of tokens as it uses the (preprocessed) tweet string as input. We further processed these vectors using the umap dimensionality reduction alg to get a vector with 5-10 dims for each tweet. This number of vectors has been found in other contexts to retain sufficient information for clustering.

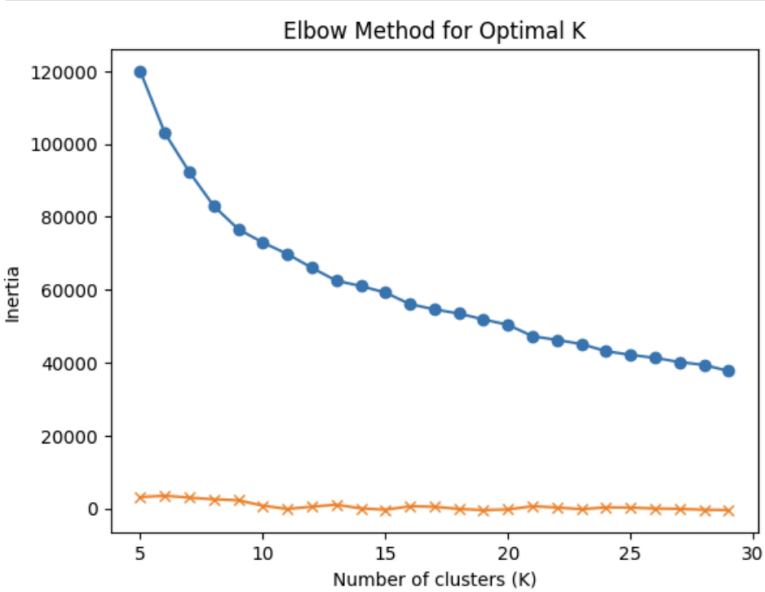
Metrics

These are the metrics used in topic modelling evaluation:

- Avg Silhouette Score
    - Avg Silhouette is a scoring of how similiar the points of each cluster are with respect to how different they are to points of other clusters.
    - Interpretation: Best = +1, Inconclusive=0, Error Warning=-1
- Variance Ratio Score
    - This score evaluates the variance within a cluster compared to the variance between cluster centroids
    - Interpretation: Best = Relative Maximum
- Davies Bouldin Score
    - Evaluates the distance of points to their cluster centroids as compared to the distance between cluster centroids
    - Interpretation: Best = 0
- Subjective Coherence & Cluster Keyword Relevance
    - Inspect the clusters manually to see if they intuitively make sense.
    - Do the top keywords for each cluster seem relevant to each other?
    - Do the keywords for each topic seem relatively distinct?

Model Evaluation

We we generally explored two categories of clustering algorithms, those that automatically choose k clusters, and those which did not. In order to get a rough idea of the best k values to choose, we evaluated k-means using the elbow method for the inertia of the clusters.



We found k=9 and k=24 to be good choices with respect to local maxima of the second derivative of the inertia(k) line.

Based on these k values, we compared the models: K Means, LDA and HDBScan. We did a preliminary investigation of Agglomerative clustering but found the runtime to be prohibitive and the performance to be relatively low.

Here are the metrics for the runs of the models we evaluated.

	Model	Variance Ratio Score	Davies Bouldin Score	Avg Silhouette Score
0	kmeans_9	4100.402329	1.090415	0.301343
1	kmeans_24	3207.895413	1.132438	0.284925
2	lda_9	21.439924	18.401831	-0.049958
3	lda_24	12.473309	24.563250	-0.077033
4	hdbscan_50_15	1435.049958	0.259962	0.645049

Based on these results, we found kmeans_9 and kmeans_24 to have the best metrics. Hdbscan_50_15 only generated 2 clusters with the given params which did not provide enough distinction. We further evaluated the models based on their top keywords based on TF-IDF.

Inspecting topics produced by kmeans_24

	3	8	15	4	0	7	10	21	1	13	9	12
size	733	690	617	595	574	527	483	467	459	428	422	421
1	0.086 arctic	0.023 people	0.040 science	0.068 environment	0.095 dont	0.055 real	0.049 new	0.137 trump	0.024 nasa	0.059 sea	0.048 blame	0.090 weather
2	0.078 ice	0.020 world	0.037 dont	0.064 energy	0.072 stop	0.038 action	0.035 year	0.040 republican	0.022 new	0.055 hurricane	0.044 hoax	0.086 hot
3	0.064 snow	0.018 stop	0.032 scientist	0.059 green	0.049 cant	0.037 people	0.031 week	0.039 donald	0.020 scientist	0.047 flood	0.035 like	0.067 year
4	0.053 winter	0.016 like	0.025 say	0.050 action	0.041 never	0.031 good	0.021 next	0.027 news	0.019 million	0.042 level	0.034 damn	0.047 extreme
5	0.052 cold	0.015 get	0.022 think	0.039 carbon	0.040 nothing	0.031 know	0.021 blog	0.026 president	0.018 science	0.040 rising	0.033 lie	0.042 record
6	0.044 canada	0.014 year	0.021 stop	0.039 sustainability	0.035 believe	0.030 earth	0.021 action	0.021 say	0.018 ha	0.034 water	0.026 know	0.041 heat
7	0.031 polar	0.014 one	0.020 denying	0.032 tree	0.025 wont	0.028 human	0.020 conference	0.020 administration	0.017 world	0.033 ocean	0.025 think	0.038 summer
8	0.030 melting	0.014 ha	0.018 denier	0.030 environmental	0.025 people	0.027 still	0.018 today	0.019 stop	0.017 billion	0.031 florida	0.025 would	0.037 must
9	0.029 action	0.013 need	0.017 like	0.025 nature	0.024 slow	0.027 like	0.018 day	0.019 obama	0.017 help	0.031 rise	0.024 one	0.035 reason
10	0.029 glacier	0.013 earth	0.017 deny	0.023 help	0.023 still	0.025 thought	0.018 world	0.019 american	0.016 trump	0.030 drought	0.024 people	0.035 hottest
11	0.026 bear	0.012 make	0.016 people	0.022 forest	0.021 deny	0.024 made	0.017 report	0.018 ha	0.015 like	0.029 flooding	0.023 real	0.035 youre
12	0.024 iceberg	0.012 much	0.016 denial	0.021 learn	0.021 one	0.024 believe	0.016 impact	0.018 report	0.015 government	0.027 due	0.023 made	0.034 please
13	0.022 canadian	0.012 time	0.015 one	0.021 solar	0.021 ha	0.024 time	0.016 video	0.018 bill	0.015 people	0.025 fire	0.022 may	0.032 last
14	0.020 age	0.012 problem	0.015 evidence	0.021 water	0.019 like	0.023 think	0.015 join	0.017 issue	0.014 make	0.022 wildfire	0.022 bad	0.031 degree
15	0.020 year	0.012 war	0.015 scientific	0.020 world	0.019 world	0.022 would	0.015 twitter	0.017 hoax	0.014 think	0.021 heat	0.022 wrong	0.030 day

Inspecting topics produced by kmeans_9

	1	5	2	6	7	0	3	8	4
size	2236	1784	1351	1332	877	808	792	721	99
1	0.032 trump	0.036 new	0.036 real	0.034 energy	0.059 latest	0.078 dont	0.080 arctic	0.031 world	1.356 thing
2	0.017 people	0.027 science	0.032 action	0.034 environment	0.056 thanks	0.055 stop	0.072 ice	0.029 blame	1.352 style
3	0.016 stop	0.024 scientist	0.024 people	0.033 action	0.049 daily	0.040 cant	0.061 snow	0.027 like	1.211 geography
4	0.016 world	0.023 obama	0.022 need	0.028 water	0.049 year	0.030 never	0.054 winter	0.027 hoax	1.135 thread
5	0.015 ha	0.021 fight	0.018 believe	0.027 green	0.047 hot	0.030 nothing	0.050 cold	0.025 would	0.554 people
6	0.015 like	0.020 paris	0.018 know	0.026 help	0.043 weather	0.028 believe	0.041 canada	0.024 people	0.000 blunt
7	0.014 say	0.019 join	0.018 like	0.024 carbon	0.026 today	0.023 still	0.029 polar	0.022 say	0.000 triage
8	0.013 get	0.018 action	0.018 say	0.023 food	0.026 day	0.023 ha	0.028 action	0.021 bad	0.000 reuse
9	0.013 science	0.017 news	0.018 help	0.023 forest	0.023 extreme	0.022 one	0.027 melting	0.021 worse	0.000 geared
10	0.013 scientist	0.016 pope	0.017 human	0.020 tree	0.022 new	0.020 deny	0.027 glacier	0.021 could	0.000 bran
11	0.013 news	0.016 talk	0.017 earth	0.020 sea	0.022 love	0.019 people	0.025 bear	0.020 damn	0.000 hovering
12	0.013 think	0.016 say	0.017 de	0.020 ocean	0.022 record	0.019 wont	0.023 year	0.020 vulnerable	0.000 akong
13	0.012 one	0.015 world	0.017 di	0.019 environmental	0.021 heat	0.019 without	0.022 iceberg	0.020 may	0.000 salinity
14	0.012 republican	0.014 china	0.016 get	0.019 impact	0.021 time	0.019 denying	0.020 canadian	0.019 know	0.000 banksters
15	0.011 dont	0.014 work	0.016 good	0.019 pollution	0.020 reason	0.018 action	0.019 weather	0.018 lie	0.000 snl

Subjective inspection of the clustering reveals most of the keyword groups to have a sense of cohesion with some outliers such as k_means group 4. Some top cluster labels could be 'Conservative Politics', 'Liberal Politics', 'Environment', 'Science', 'Ice Impact'. Based on these clusters, either kmeans_9 or kmeans_24 would be reasonable.

Final Model Choice

We select the **kmeans_9** model run with umap reduced BERT Sentence embeddings as our top performing model based on its cohesive topics and top relative metrics. The topics produced were relatively cohesive and can be used to highlight areas of importance in Climate Change public conversation to inform decision making stakeholders.