How to Spot a Climate Change Skeptic

Julia Taussig

Overview: The Data Science Process

- Problem Statement
- Data Collection
- Data Cleaning & Exploratory Data Analysis (EDA)
- Preprocessing & Modeling
- Model Evaluation
- Conclusion and Recommendations

Problem Statement

Goals:

- Use Reddit's API to collect posts from the following subreddits:

  r/climate: "a community for truthful science-based news about climate and related politics and activism"

  r/climateskeptics: "questioning climate related environmentalism"

- Binary classification problem:

  Use natural language processing (NLP) to train a classifier on which subreddit a given post came from (evaluate models using accuracy of the classifiers).

Sources: https://www.reddit.com/r/climate and https://www.reddit.com/r/climateskeptics

Data Collection

Used Reddit's API along with the Python request library (and Google Chrome)

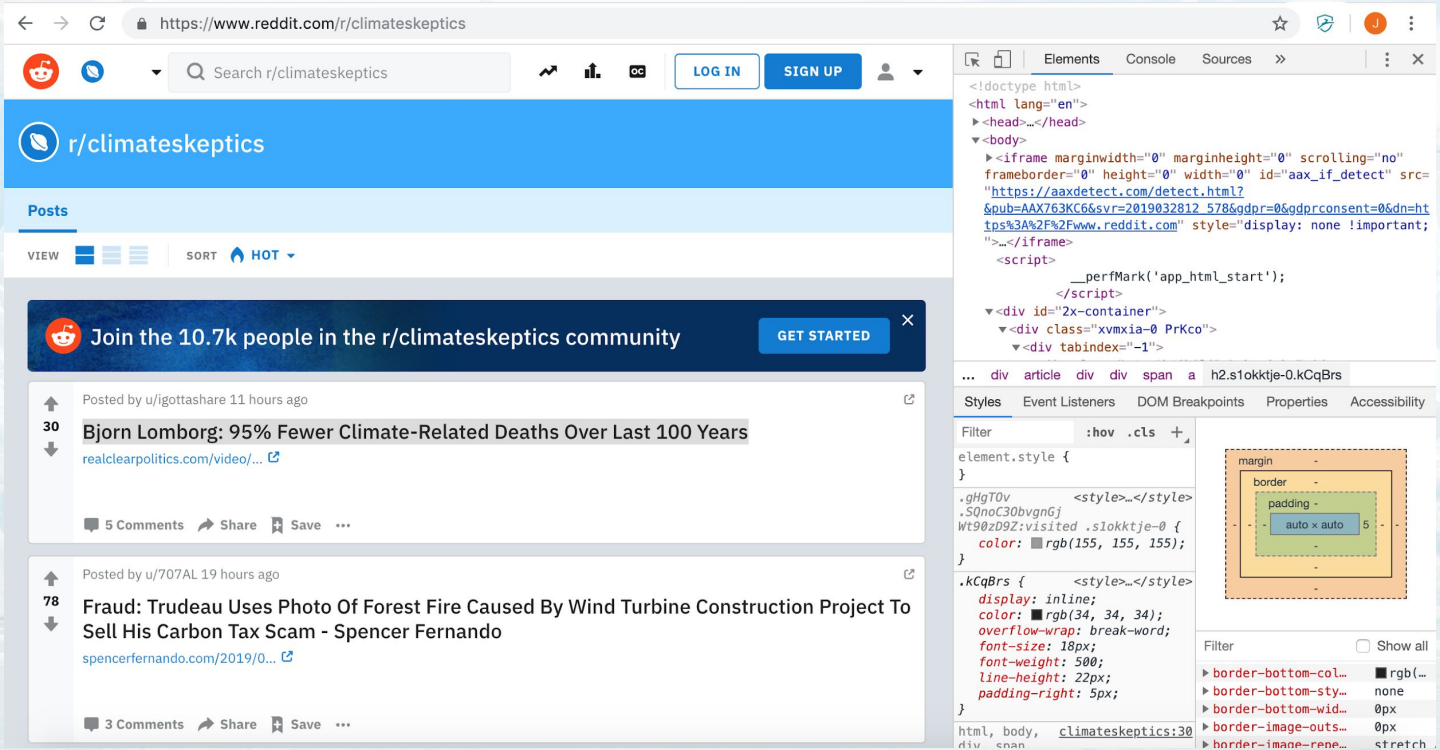


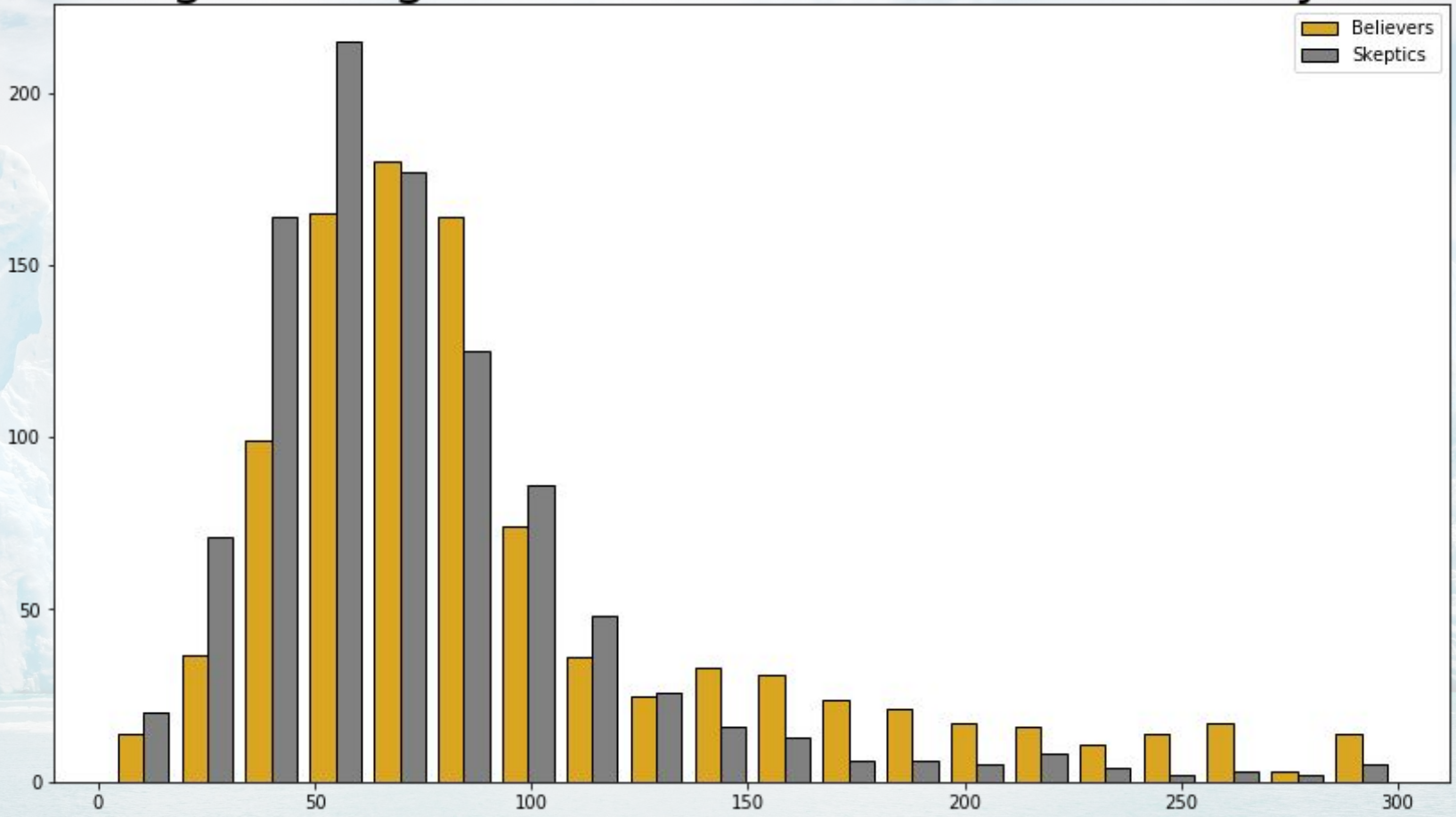
Photo by Alto Crew on Unsplash

Data Cleaning

- Data was imported into Pandas dataframes, combined into one dataframe, and inspected
- "Unnamed: 0" column was removed from each dataframe
- Reset index values
- Added target column to dataframe
    - Value of 0: post came from the r/climate subreddit
    - Value of 1: post came from the r/climateskeptics subreddit
- Null values were then inspected and removed (all in the post_text column)
- Dataframe column datatypes were inspected

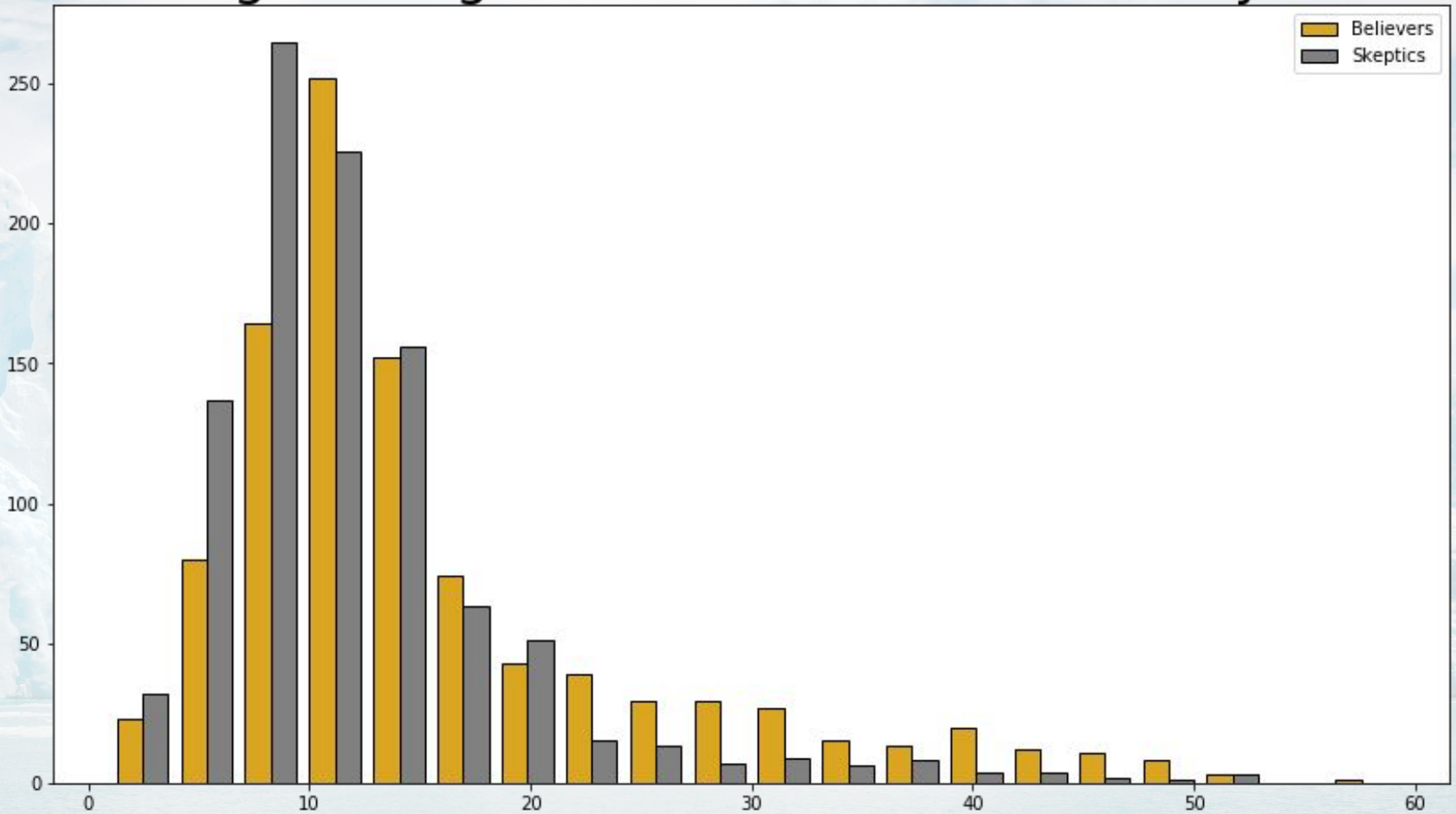
EDA

Climage Change Subreddits: Character Count by Class



EDA

Climage Change Subreddits: Word Count by Class



EDA - Generating Baseline Model

- Target variable value counts were inspected.
- 1002 r/climateskeptics posts
- 995 r/climate posts
- Approx. 50.2% of the subreddit posts from r/climateskeptics
- Approx. 49.8% of the subreddit posts from r/climate
- Good balance in the target class

Baseline model: assume any post in the test set is from r/climateskeptics

Therefore, baseline model accuracy: 50.2%

Preprocessing

- Subreddit titles were changed to all lower-case
- Punctuation was removed
- Words/terms were tokenized (split)
- Words/terms were lemmatized
- English stopwords were removed
- Post titles were joined back into strings (and added to the dataframe)
- Count vectorized or term frequency—inverse data frequency (TF-IDF) vectorized the post titles, depending on the model

EDA - Heatmap of 15 most frequently appearing words in both climate and climate skeptics post titles

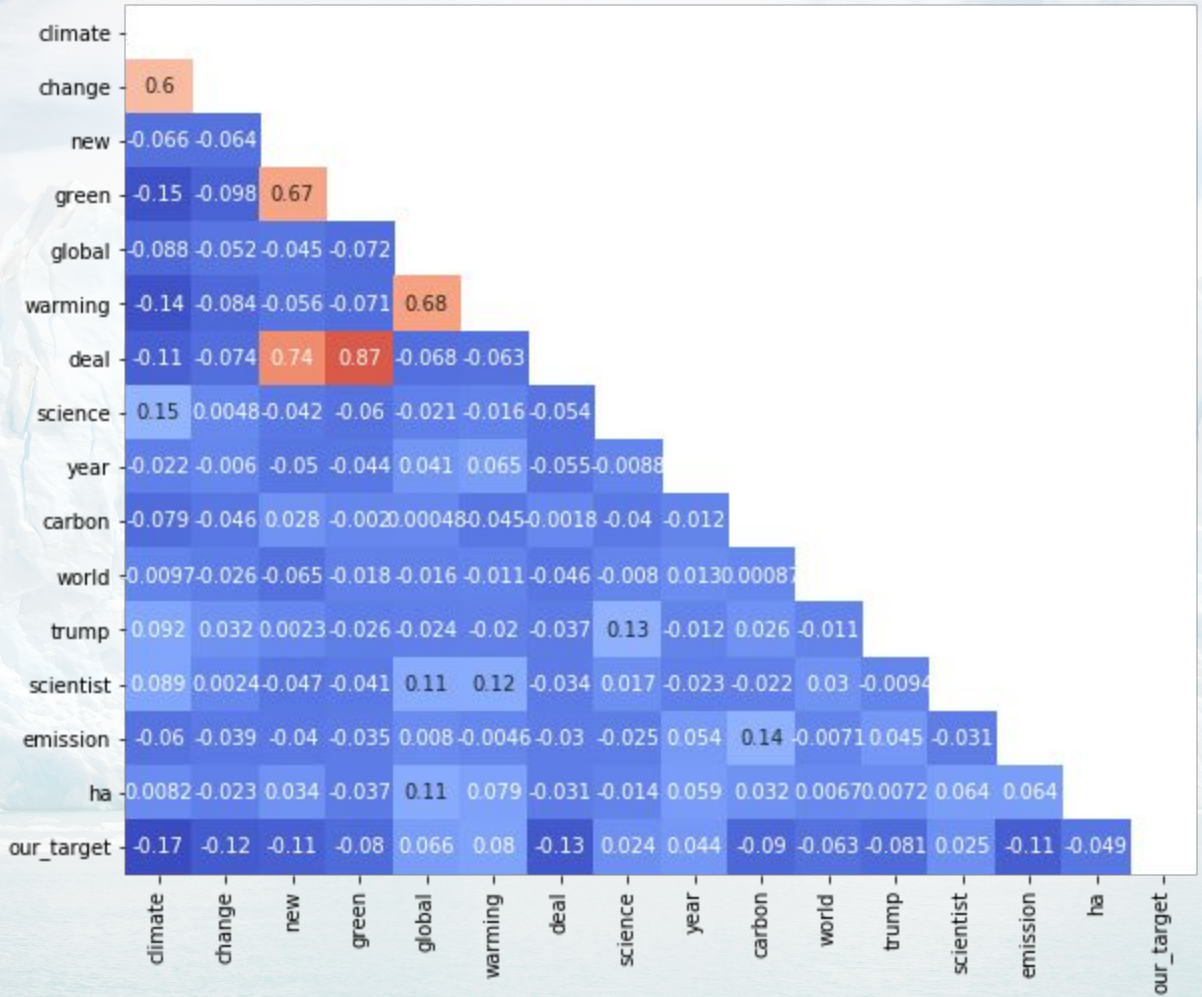
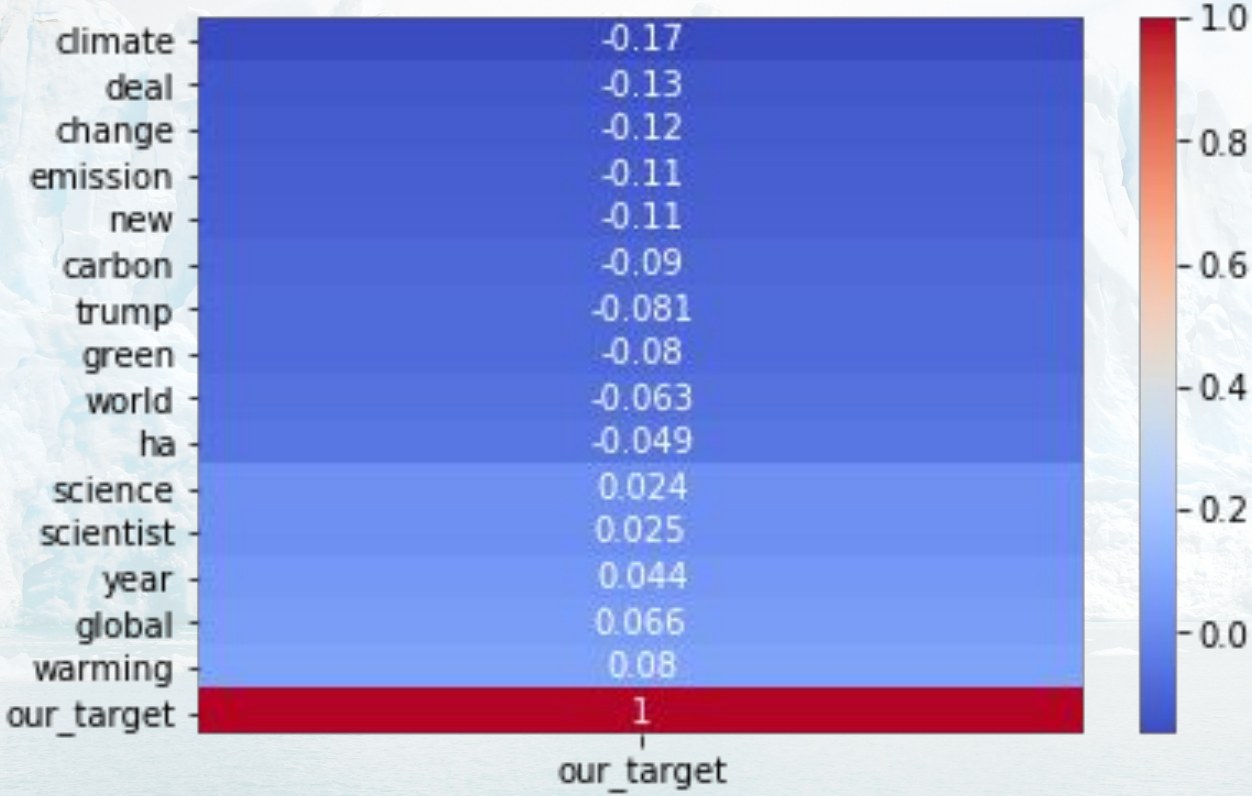


Photo by Alto Crew on Unsplash

EDA - Heatmap of 15 most frequently appearing words in both climate and climate skeptics post titles (focusing on our_target: whether from climate or climate skeptics subreddits)



EDA

Top 10 words least correlated with skeptic posts	Approx. correlation with target (skeptic post = 1)
climate	-0.166
deal	-0.133
change	-0.122
emission	-0.113
new	-0.108
plan	-0.105
action	-0.101
strike	-0.0934
carbon	-0.0899
republican	-0.0880

Top 10 words most correlated with skeptic posts	Approx. correlation with target (skeptic post = 1)
alarmist	0.114
data	0.0993
solar	0.0977
Delingpole	0.0977
alarmism	0.0807
warming	0.0803
skeptic	0.0770
NASA	0.0707
cold	0.0683
cooling	0.0682

EDA - Term Frequencies and Word Counts in Subreddit Titles

Highest term frequency and count in climate skeptic post titles

our_target	0	1
climate	0.583920	0.394212
change	0.314573	0.201597
global	0.070352	0.109780
warming	0.059296	0.105788
new	0.166834	0.091816
green	0.118593	0.069860
science	0.043216	0.053892
year	0.030151	0.047904
deal	0.118593	0.042914
scientist	0.031156	0.040918

our_target	0	1
climate	581	395
change	313	202
global	70	110
warming	59	106
new	166	92
green	118	70
science	43	54
year	30	48
deal	118	43
scientist	31	41

Highest term frequency and count in climate post titles

our_target	0	1
climate	0.583920	0.394212
change	0.314573	0.201597
new	0.166834	0.091816
deal	0.118593	0.042914
green	0.118593	0.069860
global	0.070352	0.109780
warming	0.059296	0.105788
carbon	0.057286	0.019960
emission	0.056281	0.010978
trump	0.053266	0.021956

our_target	0	1
climate	581	395
change	313	202
new	166	92
deal	118	43
green	118	70
global	70	110
warming	59	106
carbon	57	20
emission	56	11
trump	53	22

EDA

Some interesting words found only in skeptic (r/climateskeptic) subreddit post titles (not in r/climate post titles):
- colder
- fraud
- debunks
- myth
- alarmism
- alarmist
- Delingpole (likely the climate skeptic James Delingpole)
- Heller (likely the climate skeptic, Tony Heller)
- Hysteria
- Cooling
- Coldest

Note: tone of negativity and fear

Some interesting words found only in global climate change believer (r/climate) subreddit post titles (not in r/climateskeptic post titles):
- access
- chance
- better
- family
- solve
- effort
- congress
- representative
- corporate
- calling
- tackling
- infrastructure
- work
- inspired
- join

Note: tone of positivity and activism

EDA - Hypothesis Test on Top 40 Words Subreddits which Overlap

Accept null hypothesis that frequency of use of the following terms is the same for both climate change believers and skeptics subreddits:
- green
- scientist
- ice
- earth
- ha (probably a word that was so harshly lemmatized that it was cut short)
- say
- planet
- people

Reject null hypothesis that frequency of use of the following terms is the same for both climate change believers and skeptics subreddits (all occurred in climate posts more than in climate skeptics posts):
- climate
- change
- global
- warming
- new
- science
- year
- deal
- world
- time
- Trump
- carbon

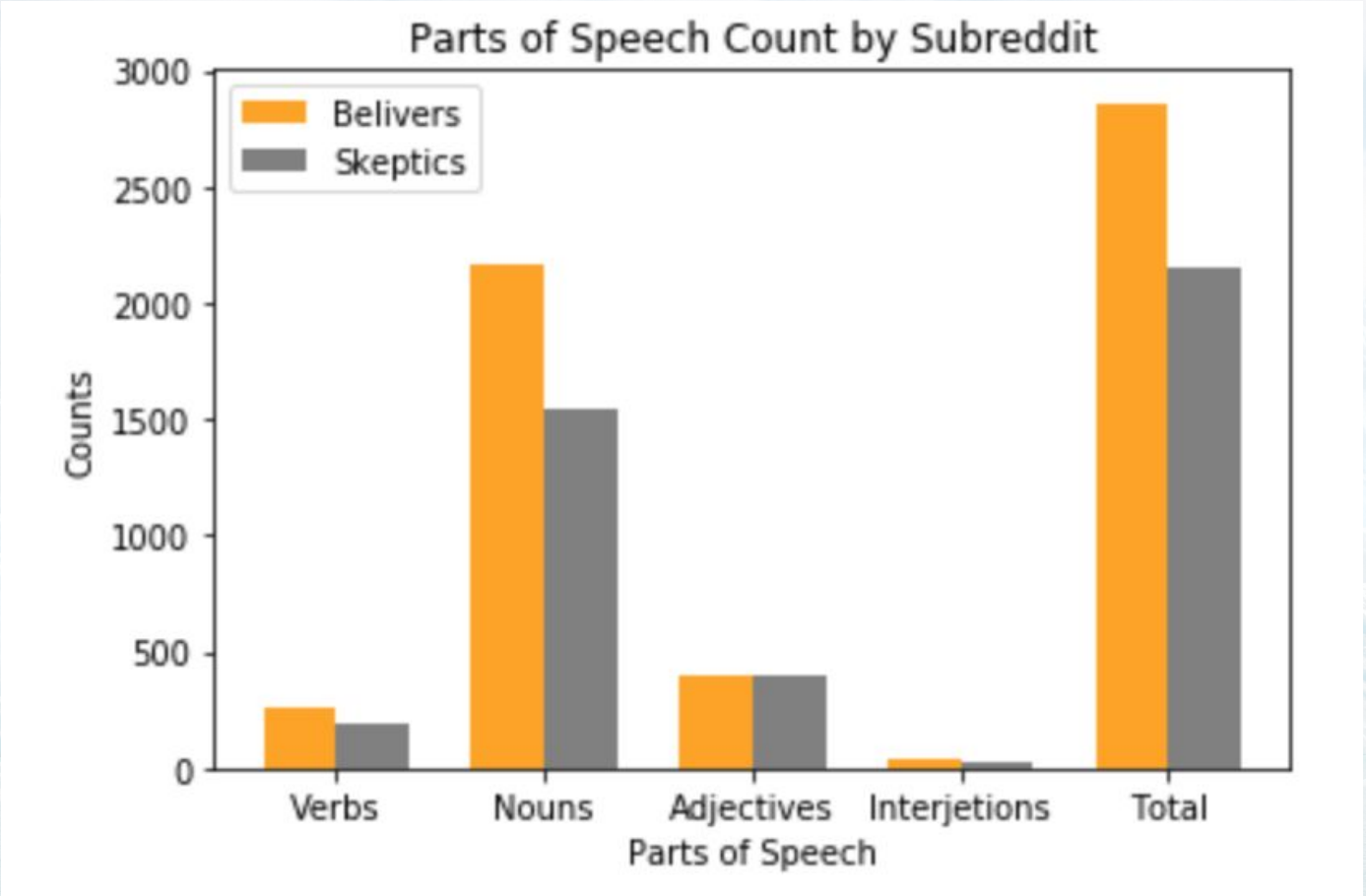
EDA - Sentiment Analysis

- Global climate change skeptic post titles slightly more negative, less neutral, and less positive than believer posts.
- The difference in sentiment was smaller than expected.

Remember: 0: r/climate, 1: r/climateskeptics

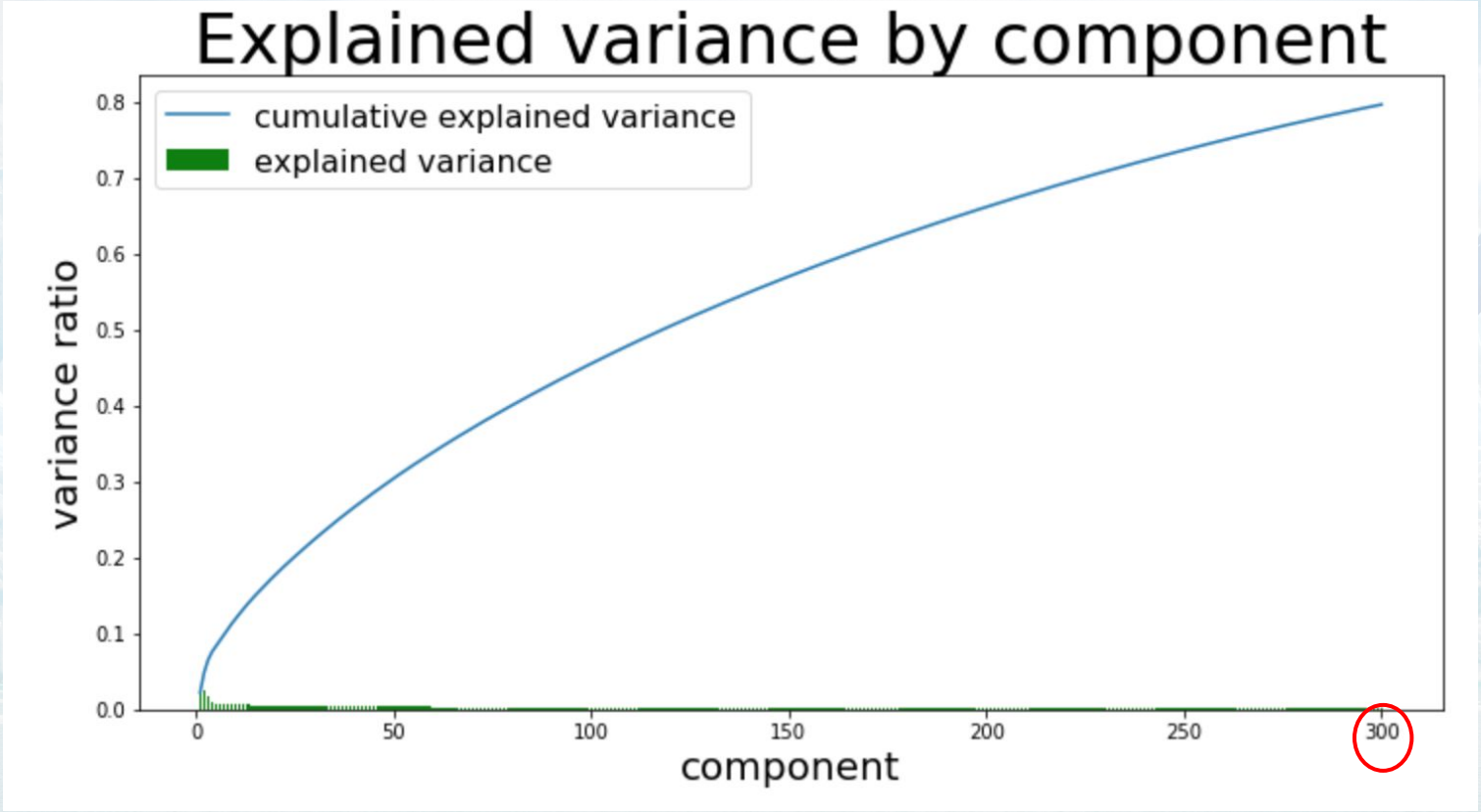
our_target	compound	neg	neu	pos
0	-0.011927	0.128767	0.745072	0.125150
1	-0.062961	0.145060	0.744445	0.110502

EDA - Parts of Speech



Modeling - Singular Value Decomposition (SVD)

Deciding on the right number of components to use for SVD (instead of computationally expensive GridSearch)



Modeling

- Count-Vectorization Parameters:
  - stop_words = 'english'
  - min_df=5
  - max_df=0.99

SVD:

- TfidfVectorizer Parameters:
  - Stop_words = ‘english’
  - min_df = 5
  - max_df = 0.99
- TruncatedSVD Parameter:
  - n_components=300

GridSearchCV Parameters:

- Regularization penalty:
  - L1: Lasso regularization
  - L2: Ridge regularization
- Inverse of regularization strength (smaller values mean stronger regularization):
  - 'C': [0.01, 0.1, 0.5, 0.7, 0.8, 0.9, 1]
- Cross-validation = 3

Modeling - Including Evaluation Using Accuracy Scores

Note: Baseline Accuracy: Approx. 0.502

Note: Models overfit (models suffer from high variance)

Model	Train Accuracy Score (Approx.)	Cross-Validation Accuracy Score (Approx.)	Test Accuracy Score (Approx.)
Naive Bayes to model count vectorized preprocessed post titles	0.769	0.657	0.676
Logistic regression on count vectorized post titles	0.841	0.677	0.664
Logistic regression on SVD-transformed post titles	0.781	0.675	0.696
Logistic regression on SVD-transformed post titles utilizing GridSearch best_params_	0.774	0.668	0.698

Conclusion and Recommendations

- We can find out whether a post is from r/climateskeptics rather than r/climate with an accuracy of approx. 67%
- Look for common skeptic words such as “global warming,” “debunk,” etc.
- Future goals:
  - Collect and analyze more data (including more post titles, post text, and post comments and features such as post authors, post popularity and emogis used and typos)
  - Optimize the train-test-split ratios
  - Try different models (and number of ngrams) on optimized count vectorized and TF-IDF vectorized and SVD-transformed data (see which parameters best to find most important features to use for the model)
  - Use other models such as boosted tree, random forest, k-nearest neighbors, decision trees, and bagged tree
  - Visualize how well model predicted values since the test target values are available for this data

Thank you!

Questions?

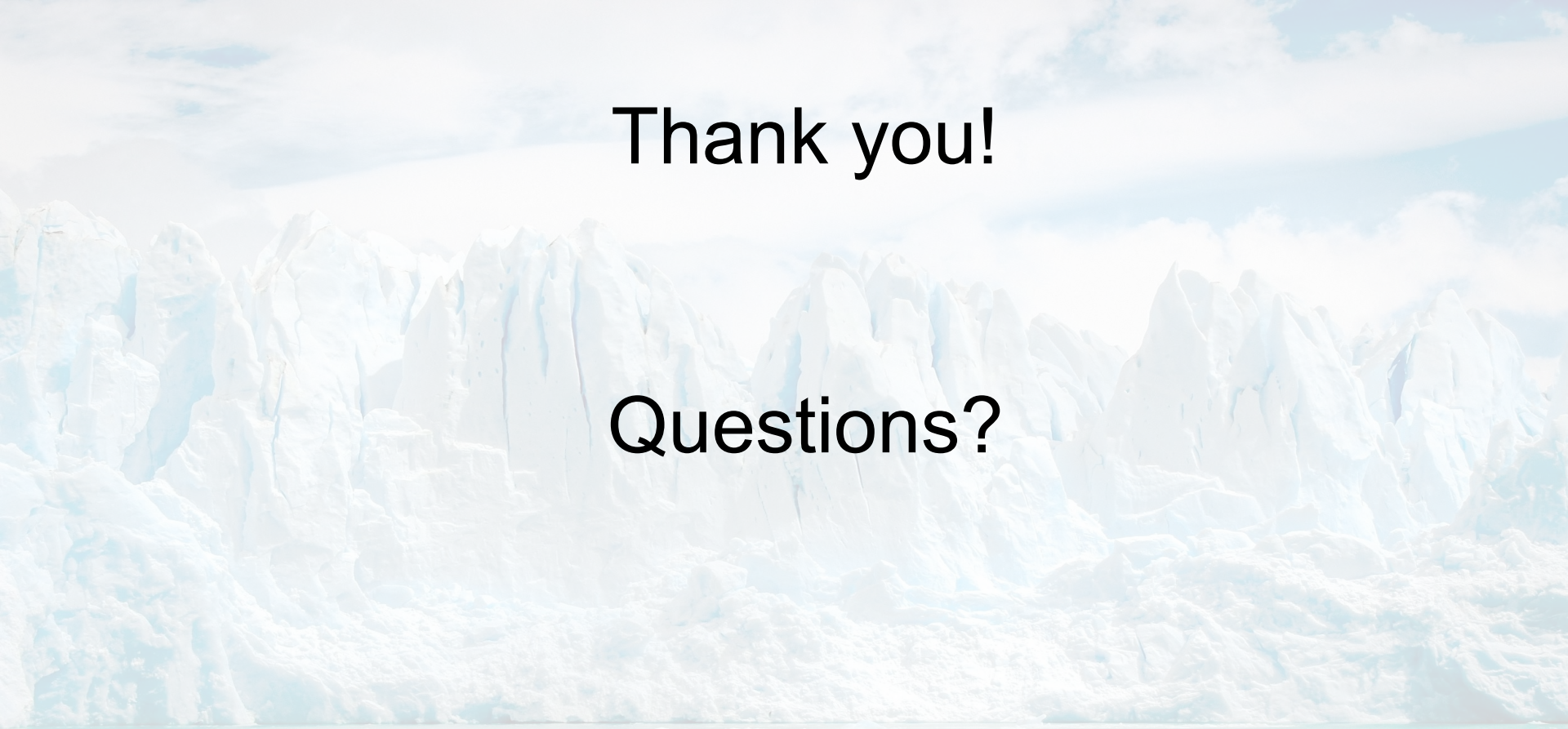


Photo by Alto Crew on Unsplash

Works Cited

https://www.reddit.com/r/climate

https://www.reddit.com/r/climateskeptics

Nagpal, Anuja, https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c

Nicol, Will, https://www.digitaltrends.com/web/what-is-reddit/

Photo by Petter Rudwall on Unsplash, https://unsplash.com/search/photos/pollution?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText

Photo by Alto Crew on Unsplash, https://unsplash.com/photos/Rv3ecImL4ak