Introduction

In the year of 2024, the unprecedented growth of social media knows no bounds, new platforms emerge daily and gain traction — all to become pertinent forms for the public sphere to thrive in, where public debate and information is available readily — especially on global issues, ranging from social media justice on poverty, social class, and climate change. Over the years, discourse related to climate change has been shaped by many mediums of the public sphere — newspapers, television, political leaders, public figures, academic journals and now, social media platforms. Each of these influences public perception in their own different way. A study done by Anderson (2009) showed us how ‘traditional’ media types (newspapers etc.) shape the public's perception of climate change, focusing more on the scientific uncertainty and political controversy associated with it. However, with the rise of social media in recent years, the type of discourse and its perception have been heavily altered and broadened the scope of discussion. Social media sites with public forums, such as Reddit, have given the people a place to voice out their unfiltered thoughts and opinions, with little to no regulation. It is a democratized form of platform, as described by (O'Neill et al., 2015). It is pertinent to notice this shift in discussion platforms, as important as it is to notice that the shift in platforms and voices heard have a significant impact in the conversation of climate change, which ultimately leads to policy change and community action.

Research conducted by Cody et al. (2015) shows us that different perceptions and narratives lead to different sentiment and polarities on social media platforms (conducted by case studies on Twitter and Facebook) and how these scores show correlation with public engagement —

leading to the conclusion that different types of topics within a sub-topic — in this case, climate change and its subtopics or “subreddits” resonate differently with various communities. Contextually, Reddit is a public forum social media platform with ‘subreddits’, akin to subtopics or subcultures of communities where individuals can post their opinions via a personal post or a comment, all of which are public. With that noted, it is important to remember that few studies have intrinsically elaborated on why this dynamic exists in the digital sphere, especially on open ended platforms such as Reddit. This research paper aims to fill that gap by providing valuable insights into how topics pertaining to climate change resonate within Reddit, a relatively unexplored platform. The paper will do so by exploring different types of natural language processing techniques such as topic modeling and sentiment analysis. The primary goal is to understand the sentiment and subjectivity of discussions around climate change on social media platforms, specifically focusing on how these discussions reflect public opinion and perceptions, and to explore the differences in sentiment and engagement within discussions about climate change across various subreddit communities. To see whether certain subreddits exhibit more positive or negative sentiments and how the engagement might be associated with these sentiments and the specific topics discussed.

Literature Review

Studies such as Jang and Hart (2015) tell us how polarization in climate change discourse changes across geographical landscapes and political divisions, something that could be seen in the subreddits we are analyzing today. Kirilenko and Stepchenkova (2014) also harp on a similar analysis by tracking discourse on climate change over the span of one year. O'Neill et al. (2015)

also looks at how public perception of climate change is altered by both ‘traditional’ (newspapers etc) media and social media. Pearce et al. (2014) looks at how Twitter can serve as a platform for discussions to measure sentiment and engagement about climate action. Williams et al. (2015) also looks at the pertinent value of understanding the impact of public forums such as social media websites, while acknowledging that they may create echo chambers, but it also tells us about more ways to approach the topic that could lead to more diversity within the realm of climate change discussion. The question then becomes — do we need more diversity or is the digital sphere populated enough with opinions?

Data and Methods

Data

This dataset contains all the posts and comments on Reddit mentioning the terms "climate" and "change", all the way until 2022-09-01. The dataset encompasses 10 variables -- including, type, id, subreddit.id, subreddit.name, subreddit.nsfw, created_utc, permalink, body, sentiment and score. It contains 22 columns.

type: Categorical variable, indicates the type of post or comment.

id: Numerical variable, a unique identifier for each entry.

subreddit.id: Numerical variable, a unique identifier for each subreddit

subreddit.name: Categorical, named category.

subreddit.nsfw: Categorical, indicating whether the subreddit is Not Safe For Work.

created_utc: Timestamp (numerical)

permalink: a URL or unique link to the comment or post.

body: Contains the text of the comment.

sentiment: Numeric representing a scale from very negative to very positive sentiment

score: Numerical score depending on the number of upvotes and downvotes a post/comment gets

Methods Overview

The variable ‘body’ is preprocessed to remove punctuation, URLs, and common stop words. Each comment is then tokenized into individual words, and further processed to identify and group bigrams and trigrams, which are sequences of two and three words that appear frequently together. Then I use the LDA model to perform topic modeling. This helps in identifying the underlying themes within the comments (done manually). Then I conducted sentiment analysis on each comment using TextBlob, which provides a measure of both subjectivity (ranging from objective facts to subjective opinions) and polarity (ranging from negative to positive sentiment). I aim to add more plots to visualize the relationship between sentiment scores and engagement scores within each subreddit, and to see which topics are associated with higher engagement and to explore whether sentiment drives this engagement, I will group my data by the assigned topic, calculating average scores for each topic. Then use ANOVA (analysis of variance testing) to determine if there are statistically significant differences in engagement scores across topics.

Preprocessing

For data preprocessing, I first manually looked at the data and ensured there were no null values, which there were none of. Then I entered the process of text normalization, which included making all text data lowercase to ensure uniformity in the text and data, removing rext elements

like special characters, tokenization, for this -- I imported python libraries nltk and spacy, which included breaking down the text into each word (‘tokens’), removing stop words (common words) such as “the” or “and” because the presence of them would be very common and possibly hinder topic results.

Topic Modeling

I first created a dictionary that would help map words and their ‘integer ids’ which is created from the processed data, then created a corpus which is essentially a list of vectors and each vector would represent a document. The vector consists of a pair of, for eg, words in the document. Then, I initialized the LDA model with multiple parameters set in it such as:

- num_topics: The number of topics to extract.
- random_state: A seed for reproducibility.
- update_every: Determines how often the model parameters should be updated.
- chunksize: The number of documents to use in each training chunk.
- passes: The total number of passes through the corpus during training.
- alpha and eta (beta in other libraries): Parameters that affect sparsity of the topics.

After the entire model is done parsing through the data and is trained, I visualize the model. Then, I extracted the topics manually to give them all names. I then inputted them into a column in the dataset called ‘topics’

Sentiment Analysis

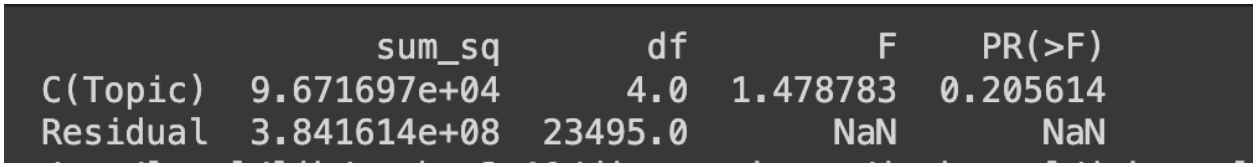
I first imported the TextBlob library to conduct sentiment analysis, trying to calculate the subjectivity and the polarity of the column ‘body’ within the dataset, which prescribes to

different posts and comments made by users on Reddit. Subjectivity measures the amount of personal opinion and factual information contained in the text. The score ranges from 0 to 1 where 0 is very objective and 1 is very subjective. Similarly, polarity is a measure of the sentiment expressed in the text. It ranges from -1 (negative) to 1 (positive).

ANOVA

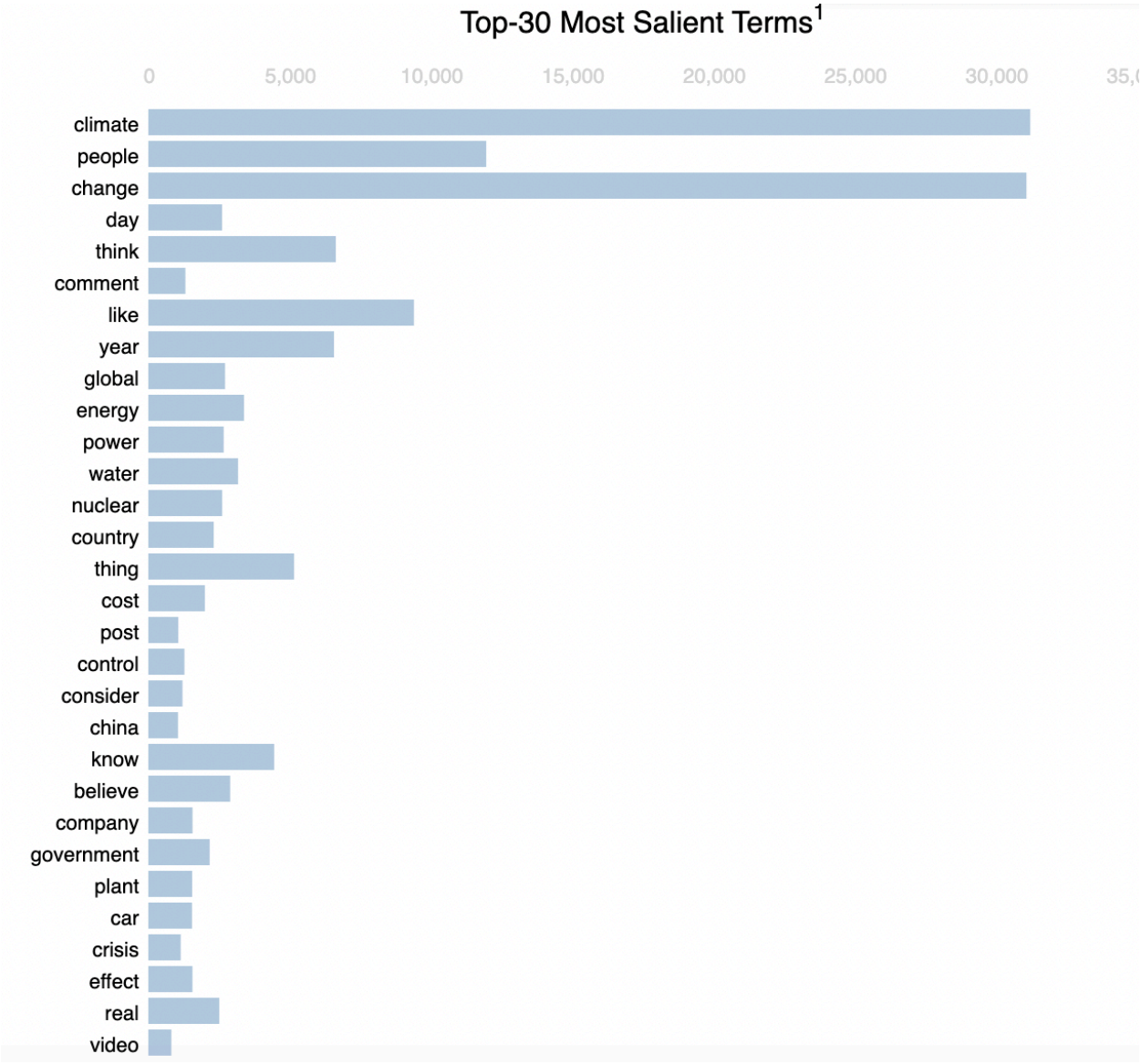
The ANOVA is significant in this context because it provides a method to test the hypothesis that different topics generate different levels of engagement. With regards to this research question, the results given tell us that differences in the mean scores across topics are not significantly significant due to the high p-value in the results.

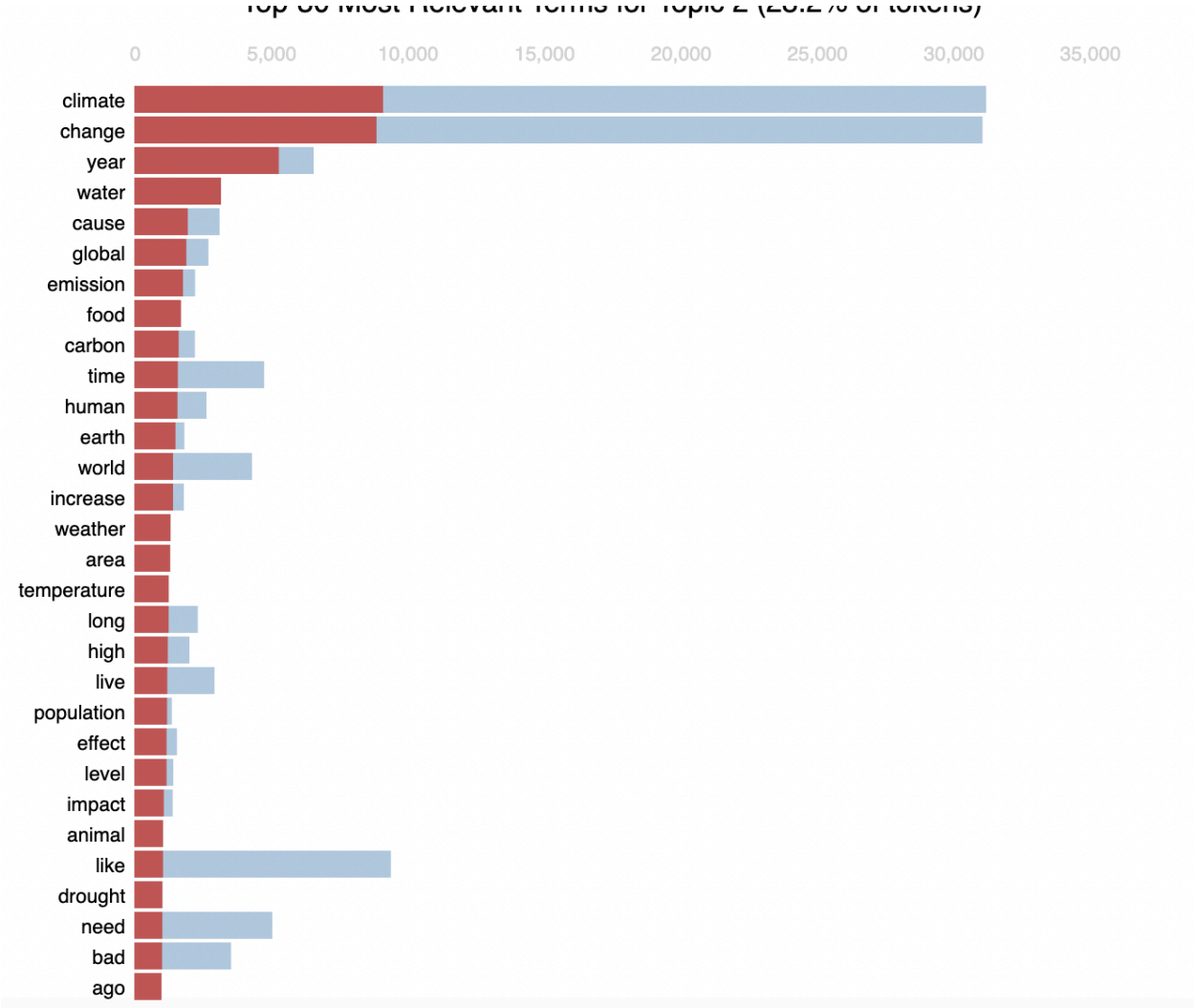
This can tell us that variability in scores in the dataset may not have anything to do with topics.

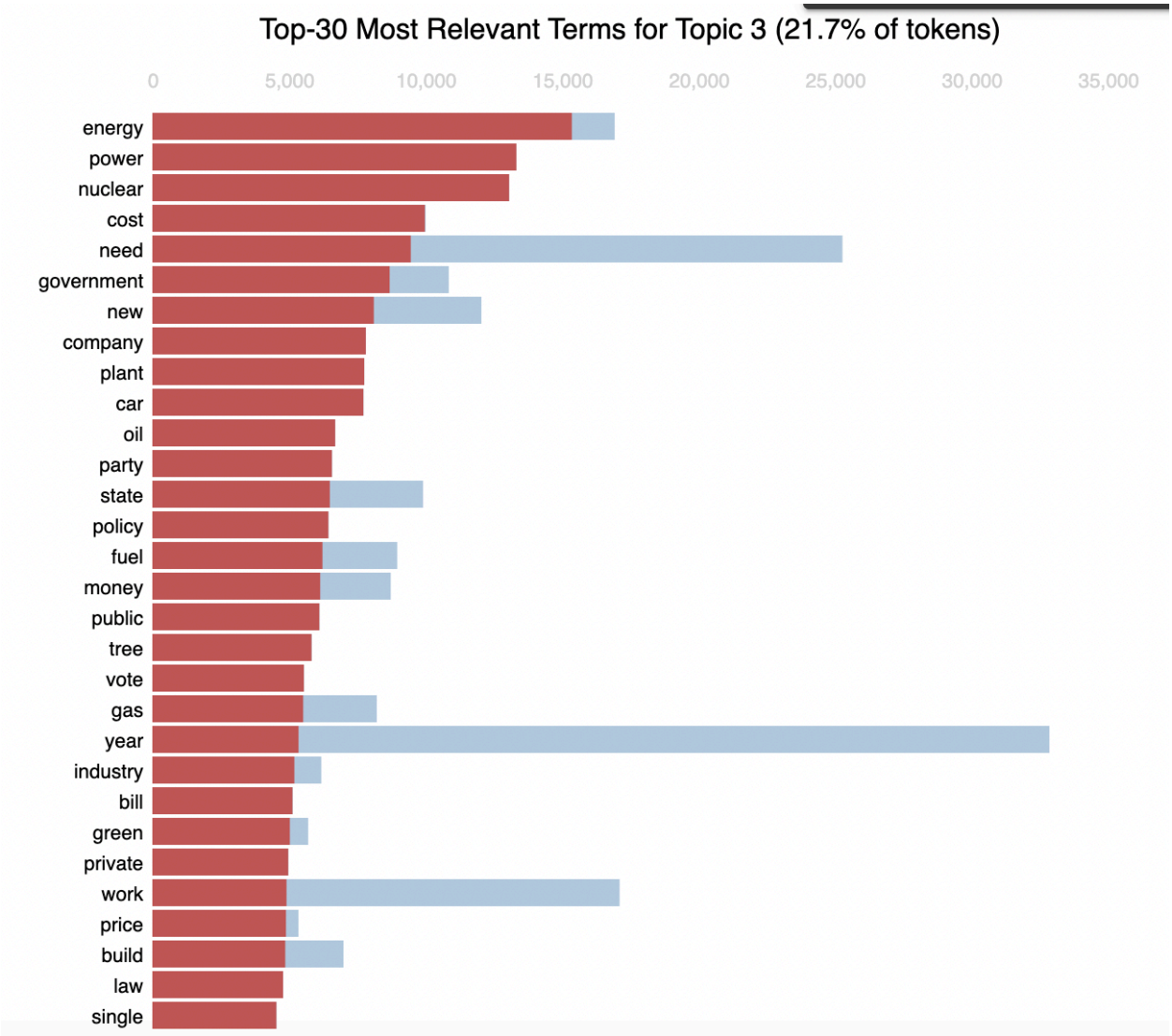


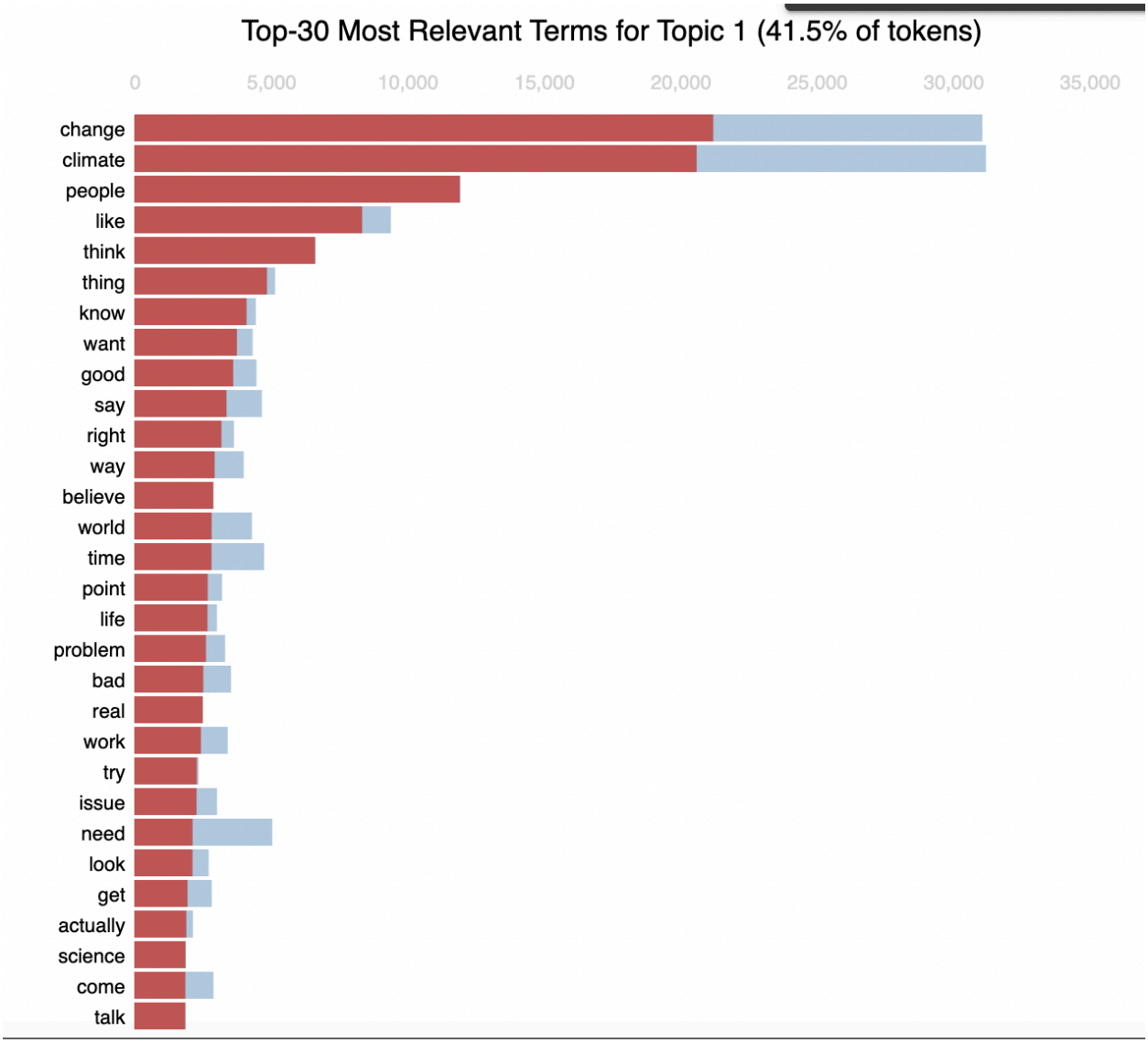
	sum_sq	df	F	PR(>F)
C(Topic)	9.671697e+04	4.0	1.478783	0.205614
Residual	3.841614e+08	23495.0	NaN	NaN

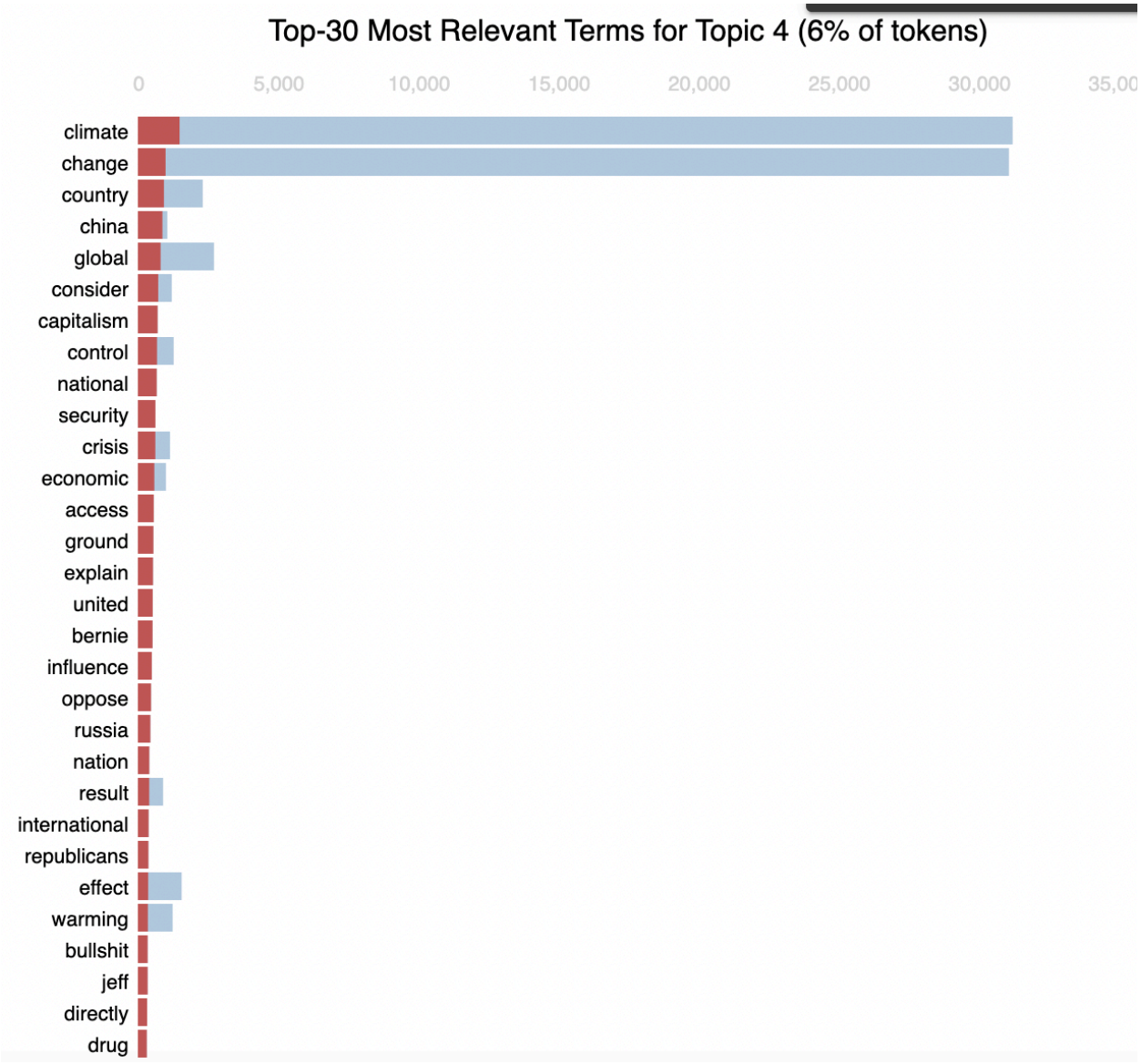
Results

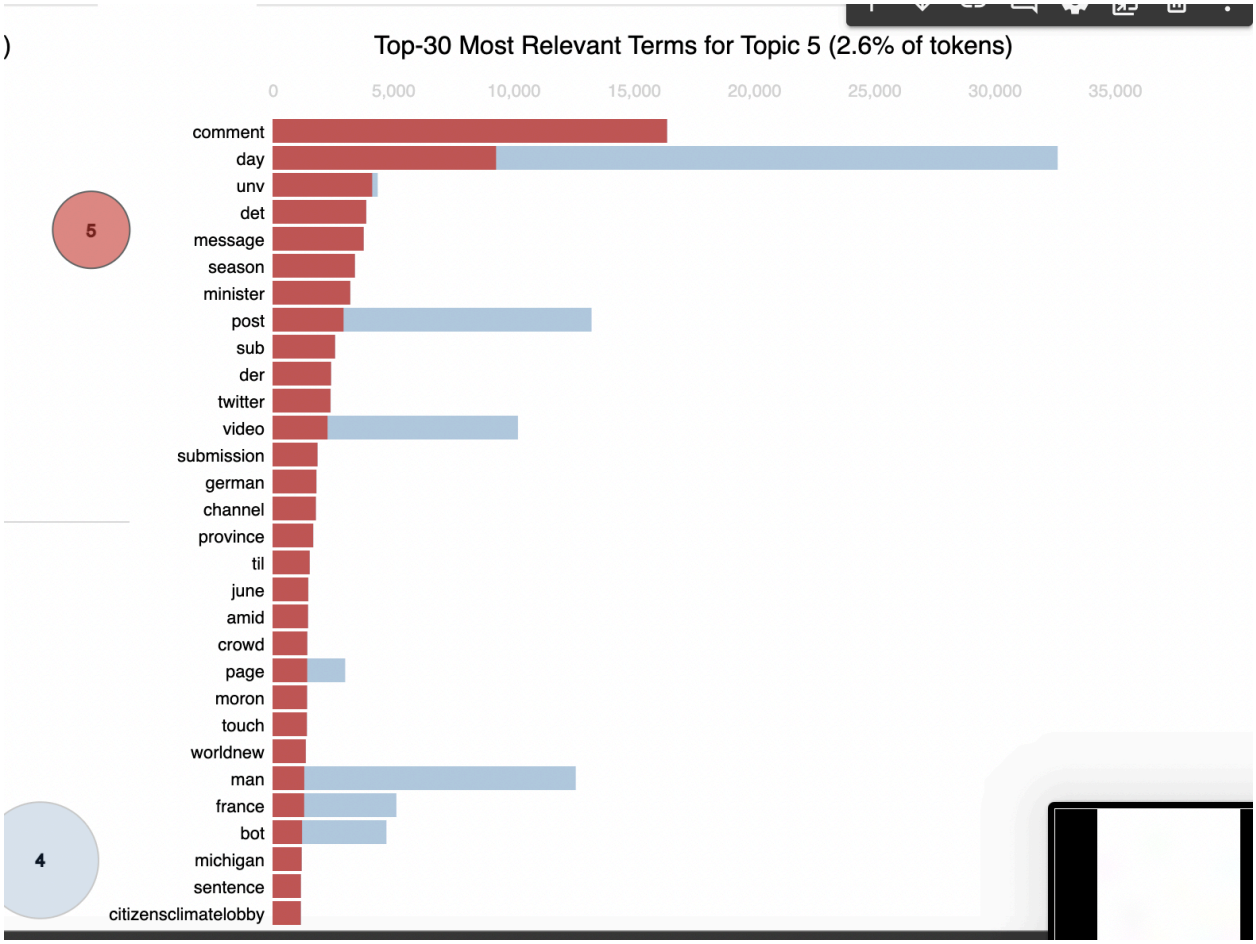

Top-30 Most Salient Terms[1]
0
5,000
10,000
15,000
20,000
25,000
30,000
climate
people
change
day
think
comment
like
year
global
energy
power
water
nuclear
country
thing
cost
post
control
consider
china
know
believe
company
government
plant
car
crisis
effect
real
video


0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
climate
change
year
water
cause
global
emission
food
carbon
time
human
earth
world
increase
weather
area
temperature
long
high
live
population
effect
level
impact
animal
like
drought
need
bad
ago


Top-30 Most Relevant Terms for Topic 3 (21.7% of tokens)
0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
energy
power
nuclear
cost
need
government
new
company
plant
car
oil
party
state
policy
fuel
money
public
tree
vote
gas
year
industry
bill
green
private
work
price
build
law
single


Top-30 Most Relevant Terms for Topic 1 (41.5% of tokens)
0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
change
climate
people
like
think
thing
know
want
good
say
right
way
believe
world
time
point
life
problem
bad
real
work
try
issue
need
look
get
actually
science
come
talk


Top-30 Most Relevant Terms for Topic 4 (6% of tokens)
0
5,000
10,000
15,000
20,000
25,000
30,000
35,00
climate
change
country
china
global
consider
capitalism
control
national
security
crisis
economic
access
ground
explain
united
bernie
influence
oppose
russia
nation
result
international
republicans
effect
warming
bullshit
jeff
directly
drug



Topic 0: Nuclear Energy in the Fight Against Climate Change

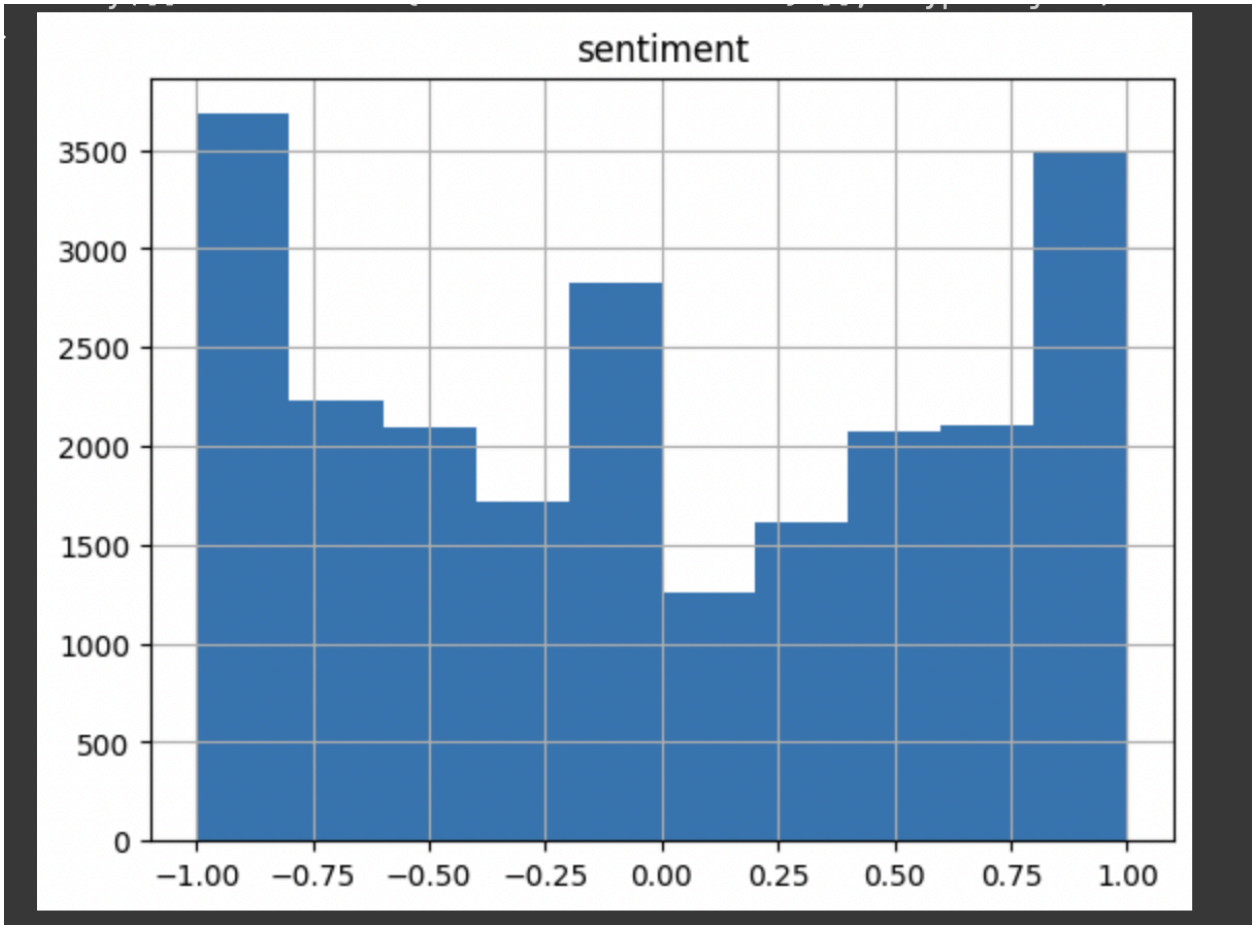
Topic 1: Public Perception and Climate Change

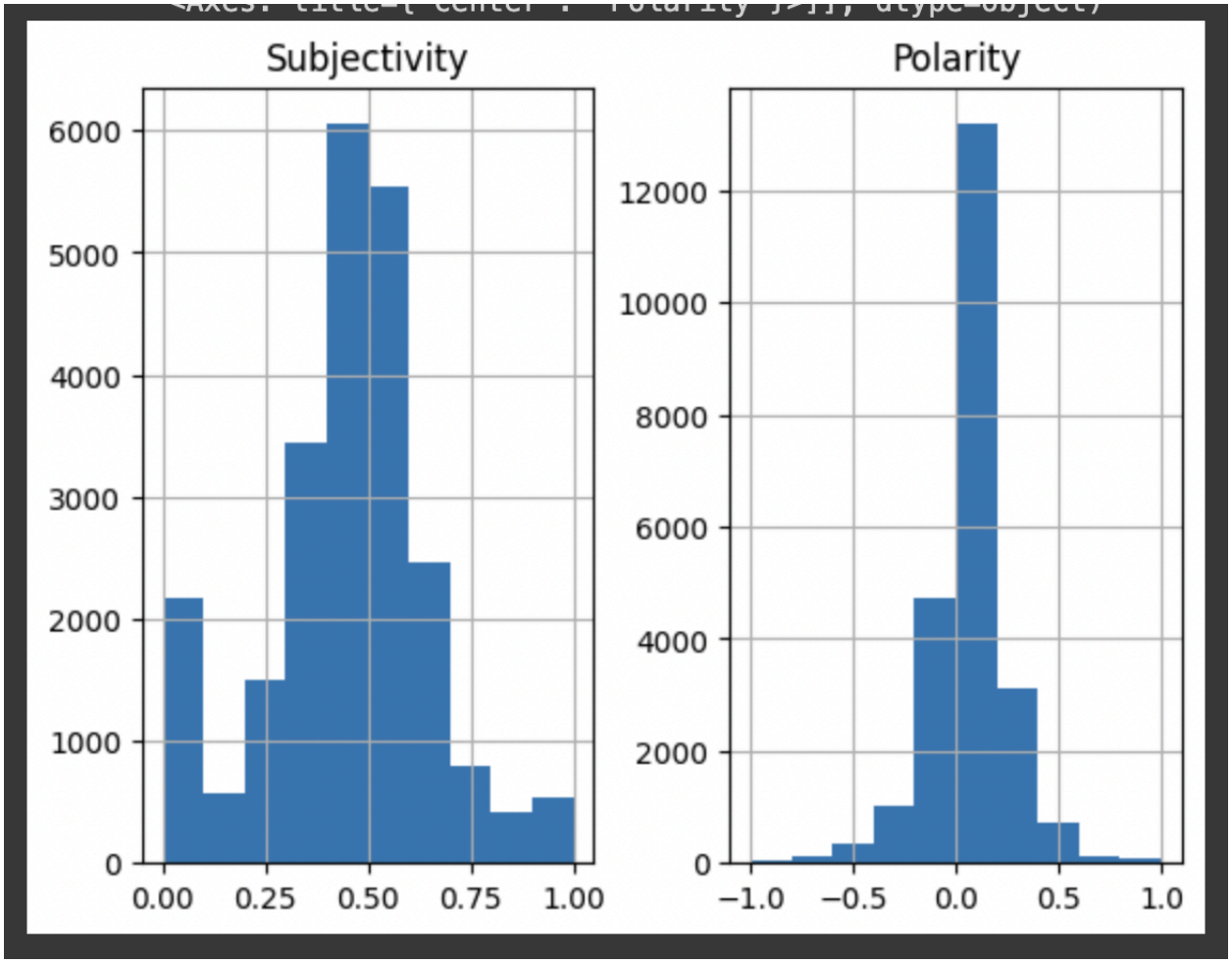
Topic 2: Long-term Effects of Climate Change on Global Water Systems and Food Security

Topic 3: Green Energy and Climate Change

Topic 4: The Geopolitical Implications of Climate Change: National Security, Economic Stability, and International Relations

Topic 5: The Impact of Social Media on Climate Change Advocacy





Mean:

Subjectivity score is 0.445725 - moderate level of subjectivity.

Polarity score is 0.062885 - overall sentiment is close to neutral, but still positive.

Standard Deviation:

Subjectivity: 0.207234

Polarity: it is 0.195397

Both show a moderate spread around the mean -- dataset shows some variation, but is not extremely skewed.

Minimum:

Subjectivity: Lowest score is 0 -- objective posts.

Polarity: Lowest score is -1, -- there was one post with a likely high negative sentiment.

Maximum:

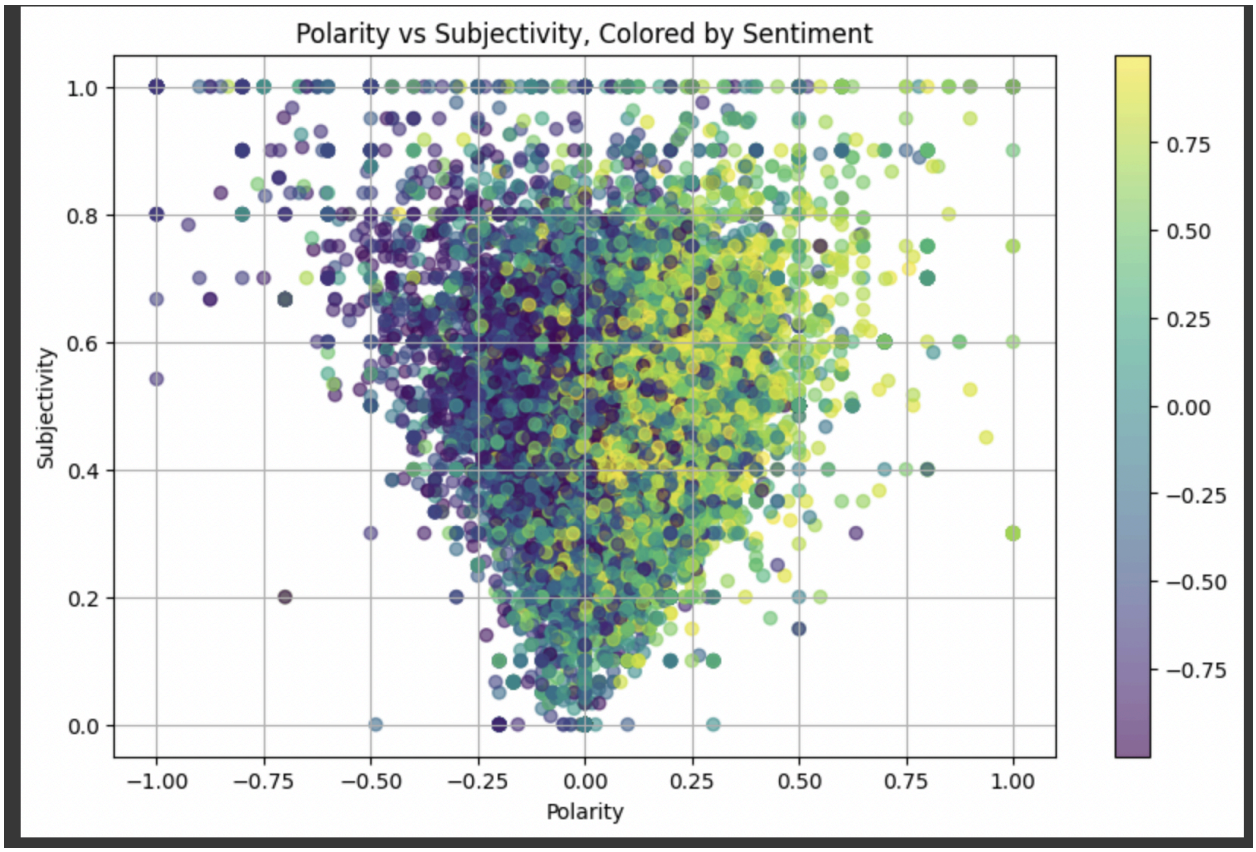
Subjectivity: Highest score is 1 - a post that was entirely subjective is likely present in the data

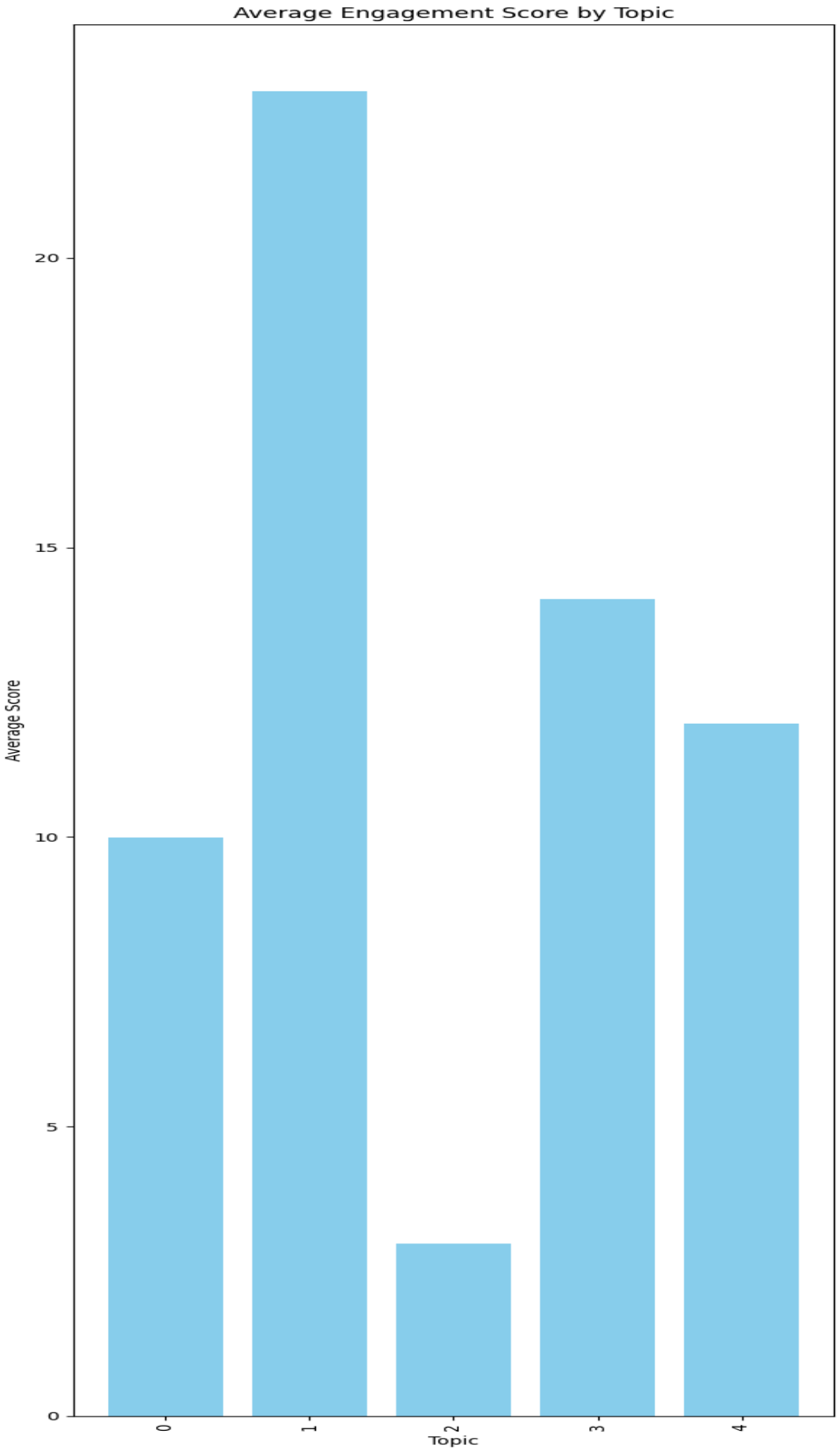
Polarity: Highest score is 1 - a post that had a very high sentiment score was present in the data.

Interpretation:

Subjectivity: Average is moderate, which shows us that the reddit posts are a good mix of factual content and public opinion.

Polarity: Average is slightly above neutral, with some very negative posts having a pull on the score, along with very positive ones. There is a good mix and a wide range of sentiments across the Reddit posts.




Average Engagement Score by Topic
Average Score
0
5
10
15
20
0
1
2
3
4
Topic

Sources

1. Jang, S. M., & Hart, P. S. (2015). Polarized frames on "climate change" and "global warming" across countries and states: Evidence from Twitter big data. Global Environmental Change, 32, 11-17.

2. Kirilenko, A. P., & Stepchenkova, S. O. (2014).Public microblogging on climate change: One year of Twitter worldwide. Global Environmental Change, 26, 171-182.

3. O'Neill, S., Williams, H. T., Kurz, T., Wiersma, B., & Boykoff, M. (2015). Dominant frames in legacy and social media coverage of the IPCC Fifth Assessment Report. Nature Climate Change, 5(4), 380-385.

4. Cody, E. M., Reagan, A. J., Mitchell, L., Dodds, P. S., & Danforth, C. M. (2015). Climate change sentiment on Twitter: An unsolicited public opinion poll. PLOS ONE, 10(8), e0136092.

5. Pearce, W., Holmberg, K., Hellsten, I., & Nerlich, B. (2014). Climate change on Twitter: Topics, communities and conversations about the 2013 IPCC Working Group 1 report. PLOS ONE, 9(4), e94785.

6. Williams, H. T., McMurray, J. R., Kurz, T., & Hugo Lambert, F. (2015). Network analysis reveals open forums and echo chambers in social media discussions of climate change. Global Environmental Change, 32, 126-138.

7. Boussalis, C., & Coan, T. G. (2016). Text-mining the signals of climate change doubt. Global Environmental Change, 36, 89-100.

8. Anderson, A. A. (2017). Effects of social media use on climate change opinion, knowledge, and behavior. Journal of Environmental Psychology, 44, 77-87.

9. Veltri, G. A., & Atanasova, D. (2017). Climate change on Twitter: Content, media ecology and information sharing behavior. Public Understanding of Science, 26(6), 721-737.

10. Lörcher, I., & Taddicken, M. (2017). Debating climate change on Facebook: A discourse analysis of community practices. Journal of Environmental Psychology, 53, 104-115.

11. Scruggs, L., & Benegal, S. (2012). Declining public concern about climate change: Can we blame the great recession? Global Environmental Change, 22(2), 505-515.

12. Painter, J., & Ashe, T. (2012). Cross-national comparison of the presence of climate skepticism in the print media in six countries, 2007-10. Environmental Research Letters, 7(4), 044005.

13. Risbey, J. S. (2008). The new climate discourse: Alarmist or alarming? Global Environmental Change, 18(1), 26-37.

14. Nisbet, M. C., & Kotcher, J. E. (2009). A two-step flow of influence? Opinion-leader campaigns on climate change. Science Communication, 30(3), 328-354.

15. Leiserowitz, A., Maibach, E., Roser-Renouf, C., Rosenthal, S., & Cutler, M. (2017).** Climate change in the American mind: October 2017. Yale University and George Mason University. New Haven, CT: Yale Program on Climate Change Communication.