Analysing and Tracking Sentiment and Topics on Social Media

Su Myat Noe Yee (s3913797)

Table of Contents

1. Introduction

The “environment” subreddit is a virtual space where individuals discuss various environmental topics, ranging from conservation to sustainability. In this analysis, we will explore the conversations within this subreddit to better understand what people are talking about and how they feel about environmental issues. The main questions for this analysis are:

- What topics are trending and popular within the “environment” subreddit?
- How do people express their feelings and opinions about environmental matters in their posts and comments?

Data are collected from subreddit “environment” through reddit API & used Natural Language Processing to pre-process data. By examining content and sentiment of discussions, we hope to gain insights into how online community perceives and interacts with environmental topics.

2. Data Collection

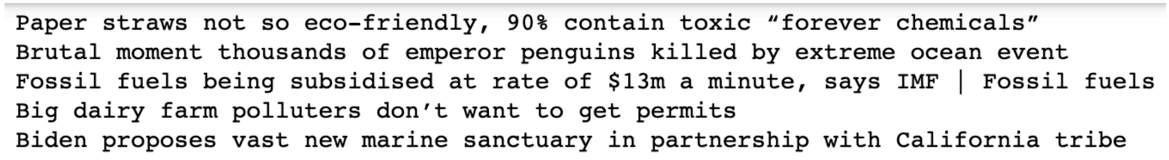
For analysis, data which is 154 hot posts were collected from “environment” subreddit on Reddit. Data collection process involved following steps:

- Reddit API Access: Access to Reddit data was through use of Reddit API which includes setting up API credentials such as client ID, client secret, password, and user agent.
- Subreddit Selection: “environment” subreddit was chosen for this analysis.
- Data Retrieval: PRAW library was used to retrieve data from the subreddit. We set 200 hot submission topics limits. However, we had to specify the end date (August 25, 2023) to make it consistent for analysis. Otherwise, every time the script is run; we will get new hot submission. After specifying date range, we are left with 154 hot submissions for analysis.
- Data Storage: The retrieved data, including titles, authors, and contents such as date, commentors, comments were stored in a structured JSON format for further analysis. This process roughly takes 120 seconds.

3. Data Exploration

Before performing text analysis and data cleaning, we will explore our subreddit on topics such as submission titles, user participation & commentors.

3.1 Titles of Hot Submissions - By iterating through hot submissions, we retrieve hot submissions titles from subreddit which provides trending topics/ discussions within subreddit.



```
Paper straws not so eco-friendly, 90% contain toxic “forever chemicals”
Brutal moment thousands of emperor penguins killed by extreme ocean event
Fossil fuels being subsidised at rate of $13m a minute, says IMF | Fossil fuels
Big dairy farm polluters don’t want to get permits
Biden proposes vast new marine sanctuary in partnership with California tribe
```

3.2 Users/Authors Participation Analysis - Next, we identified the most active authors among both hot and top submission. Figure 1 (Left) visualisation of user participation analysis revealed top 20 authors with highest submission counts.

3.3 Top Commenters in Hot Submissions – We also investigate top commenters in hot submissions by extracting usernames of commentors. Figure 2 (Right) bar chart shows engagement of top 20 commenters in subreddit's discussions.

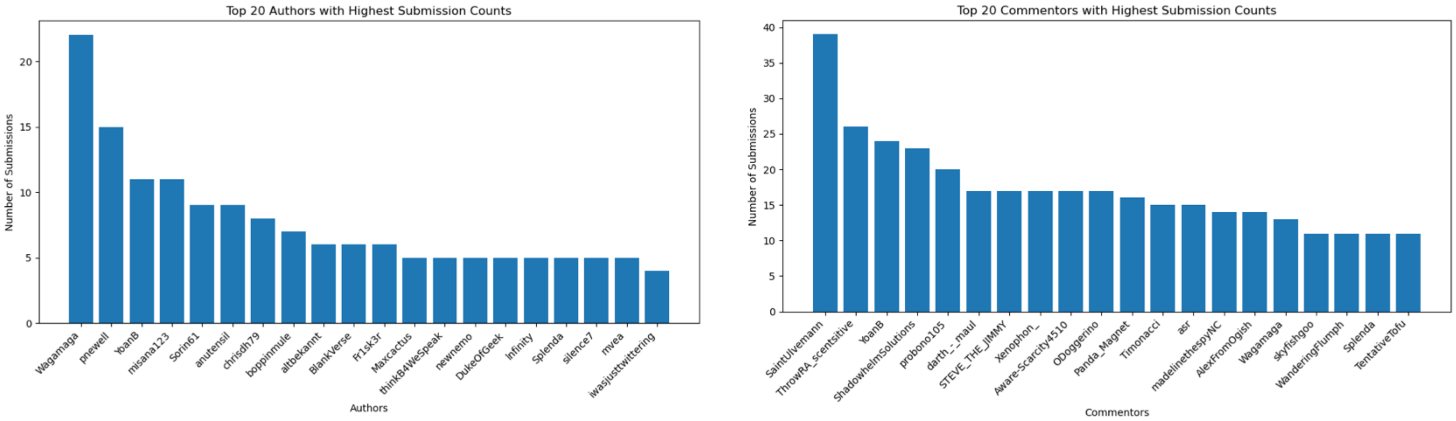


Figure 1 – Top 20 Authors and Commentors with Highest Submission Counts

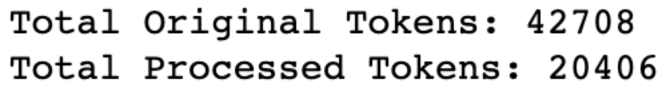
4. Text Pre-processing and Data Cleaning

Text pre-processing is a critical step in natural language processing (NLP) and text analysis. It involves transforming raw data into a structured, cleaned, and normalized format that is more suitable for analysis as it improves quality and reliability of insights gained from text data. This improves accuracy when performing text analysis algorithm: sentiment analysis and topic modelling. In this section, we apply those to submission titles and comments from subreddit.

4.1Text Pre-processing: Pre-processed text is ready for analysis after following.

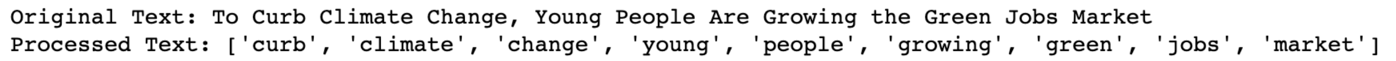
- Lowercase conversion: Transforming text data to lower letter as 'Hi' and 'hi' might be treated as two different tokens while performing text analysis.
- Tokenization: Breaks sentences into words or phrases.
- Stripping whitespaces before and after words: Inconsistent whitespaces before or after words can lead to discrepancies in analysis. 'hi' and ' hi' is treated as two different tokens if whitespace is not removed.
- Removal of stop words, digits, username non-ASCII words: Stop words are common words that does not hold much meaning such as 'a', 'an', 'the', 'and', 'have' and removing usernames which starts with @ and digits.

4.2 Tokens Before and After Data Pre-processing: Following insights shows level of noise reduction achieved through text pre-processing by calculating total number of tokens in both original and cleaned format. By doing text pre-processing, we filtered out half of noise tokens.



```
Total Original Tokens: 42708
Total Processed Tokens: 20406
```

And the following text show one of the texts from Subreddit and how it looks like after text-pre-processing. After performing text pre-processing, all the words are now in lowercase and whitespace and stop words aren't present. Moreover, the sentences are broken into words because of tokenisation for the purpose of analysing term frequency.



```
Original Text: To Curb Climate Change, Young People Are Growing the Green Jobs Market
Processed Text: ['curb', 'climate', 'change', 'young', 'people', 'growing', 'green', 'jobs', 'market']
```

4.3 Daily Count of Post: Figure 2 shows that member of community are actively participating in sharing perspectives, emotions, and opinions about environmental subject on Reddit.

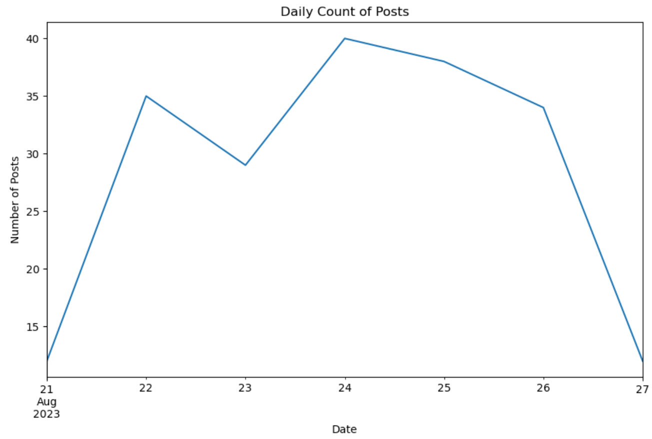


Figure 2 – Daily Count of Posts

4.4 Analysing Term Frequency in Submissions Titles vs Title and Comment: We analysed frequency of terms present in submission titles and for both submission title and comments. By processing pre-processed text data and calculating term frequency, we identified most frequent terms. Bar charts visualized top 50 most frequent terms appearing in subreddit.

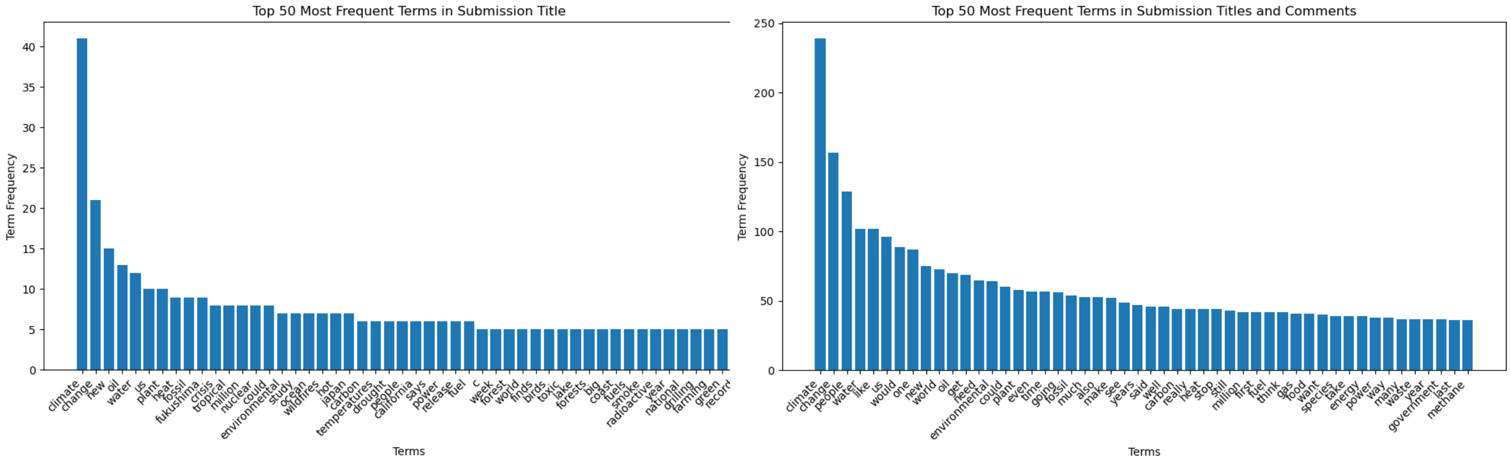


Figure 3– Top 50 Most Frequent Terms in Subreddit

5. Analysis Approach

In this section, we outline the methods and techniques used to answer the questions mentioned in Section 1 of this report. Those questions will be answered using both sentiment analysis and text analysis using topic modelling.

5.1Sentiment Analysis Approach

Sentiment analysis is the use of computation approaches such as NLP, text mining and machine learning algorithms to identify, extract and study sentiment/ opinions. It aims to determine emotional tone expressed in text data. We will be using two sentiment analysis methods for a comprehensive understanding of data (1) Opinion Word Counting Sentiment Analysis Method (2) Valence Aware Dictionary and Sentiment Reasoner (VADER) Method.

5.1.1 Opinion Word Counting Sentiment Analysis Method

Count Word method, known as Bag-of-Words method, involves counting the occurrences of positive and negative words in submission titles and comments to determine its sentiment. For this, we will use a list of positive and negative words which are stored in a text file. Finally, sentiment is determined based on which category has a higher count. The output is a sentiment label either positive, negative, or neutral based on count of positive and negative words.

5.1.2 Vader Sentiment Analysis Method

VADER is a lexicon and rule-based sentiment analysis tool designed to determine the sentiment (positive, negative, neutral). It considers both individual words in a text as well as the context in which they appear to compute sentiment score. It also accounts for punctuation, capitalization, and intensifiers to improve sentiment analysis accuracy. Afterwards, it evaluates the sentiment by analysing text and assigning positive, negative, and neutral scores to each submission title and comment. The compound score represents the overall sentiment polarity.

5.2 Text Analysis using Topic Modelling Approach

Text analysis focus content of text (submission title and comments) to extract underlying themes and topics. Latent Dirichlet Allocation method, topic modelling technique will be used.

5.2.1 Latent Dirichlet Allocation (LDA) Method

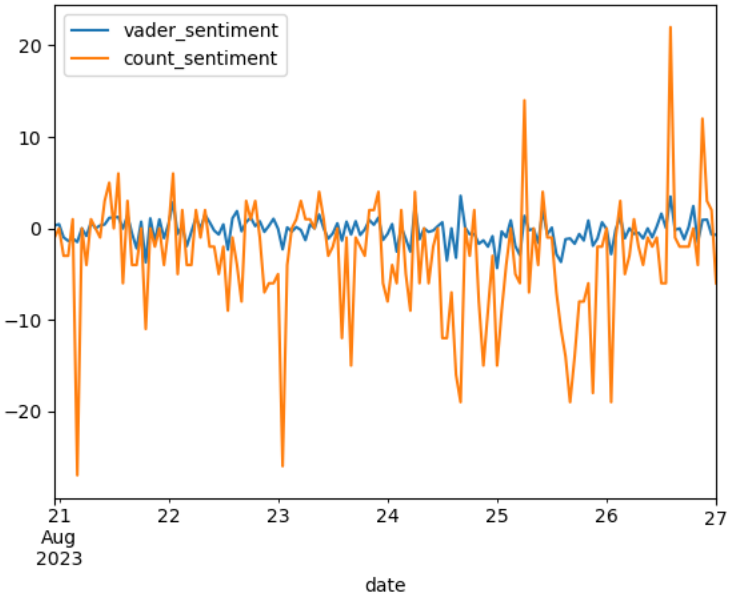
LDA is a probabilistic model that uncovers hidden topics in a collection of documents. Each submission title and comment are considered a document, and LDA identifies topics as groups of words that frequently appear together. The number of topics is pre-defined, and each topic consists of a distribution of words. We use the LDA model to extract a specified number of topics that provide insight into the dominant themes present in the Reddit data.

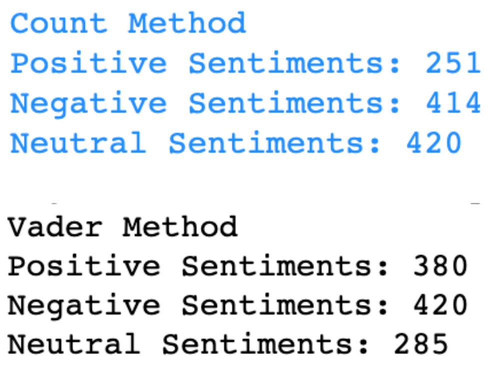
6. Analysis & Insight

This section of report will focus on the results of analysis methods employed in Section 5, aiming to uncover valuable insights from the data. We discuss outcomes of both sentiment analysis and text analysis, offering meaningful interpretations and findings.

6.1Sentiment Analysis Approach Insight

For this analysis, we performed two analyses: Opinion Word Counting Method and Vader Method as mentioned in Section 5. Left represents the classification of calculated sentiment for each text using Opinion Word Counting Method and right shows for the Vader method.





```
Count Method
Positive Sentiments: 251
Negative Sentiments: 414
Neutral Sentiments: 420

Vader Method
Positive Sentiments: 380
Negative Sentiments: 420
Neutral Sentiments: 285
```

Figure 4 – Sentiment Analysis plotted Hourly for Sentiment Analysis Method

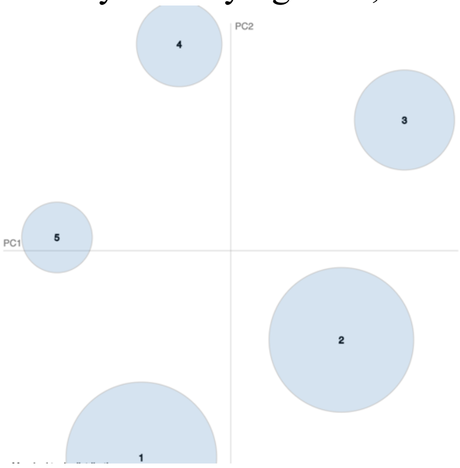
Comparing the results, "Vader Method" assigns more positive sentiments and negative sentiments and less neutral sentiments than "Count Method". This indicates that Vader's lexicon-based approach captures a wider sentiment range. Count Method identifies more neutral sentiments because it relies on specific words without context, leading to neutral assignments. This approach might lack accuracy due to ignoring word context. In contrast, Vader Method depends on sentiment lexicons, possibly not covering full language complexity.

Figure 4 illustrates sentiment analysis results using both methods over time. The y-axis shows post sentiment scores. Scores above 0.05 are positive, below -0.05 are negative, and in between are neutral. The graph displays sentiment fluctuations for both methods. There is a noticeable score differences between two approaches. This is because Count Method counts predefined positive/negative words, causing extreme scores if a text contains higher frequency of positive/negative words. However, Vader Method evaluates context and intensity, yielding nuanced scores, explaining divergent sentiment scores for the same period.

6.2 Text Analysis Using Topic Modelling Approach Insight

This section will present the insights gained from topic modelling technique, Latent Dirichlet Allocation. LDA allows us to uncover latent topics within a collection of documents, with each document representing a submission title and comment. With LDA, we can identify key themes in Reddit data. Each theme corresponds to a topic, which is a group of words that frequently appear together. These topics show us what Reddit community is discussing.

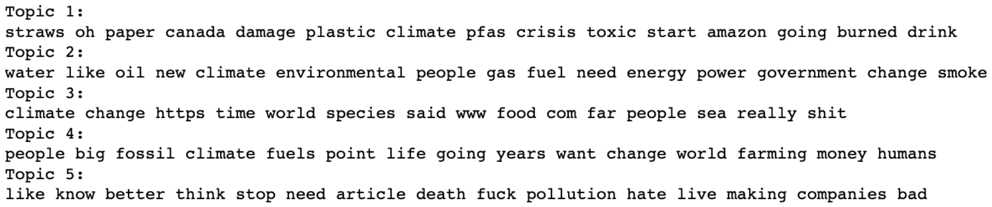
We need to define number of topics for LDA. As you can see from Figure 5, when we set number of topics to 5, the topics are not overlapping, indicating that topic modelling is successfully identifying clear, distinct, and meaningful themes within the data.



As we increase the number of topics, the total amount of tokens is divided among a larger number of topics, leading to lower percentages of tokens for each individual topic resulting in some topics having low percentages of tokens, which might indicate that those topics are not well-defined or meaningful. Thus, we decide to use 5 topics.

Figure 5 – Intertopic Distance Map with number of topics = 5

LDA analysis revealed five distinct topics that are prevalent in "environment" reddit:



```
Topic 1:
straws oh paper canada damage plastic climate pfas crisis toxic start amazon going burned drink
Topic 2:
water like oil new climate environmental people gas fuel need energy power government change smoke
Topic 3:
climate change https time world species said www food com far people sea really shit
Topic 4:
people big fossil climate fuels point life going years want change world farming money humans
Topic 5:
like know better think stop need article death fuck pollution hate live making companies bad
```

- Plastic Pollution and Its' Impact: This topic focus on the detrimental effects of plastic waste, the climate crisis, and the environmental damage caused by plastic products.
- Energy and Resource Consumption: This topic focus on concerns related to water usage, oil consumption, energy production, and need for sustainable practices.
- Climate Change and Global Impact: This topic focus on climate change's global impact, its effects on various species, and the urgent need for action.
- Fossil Fuels and Transition to Alternatives: This topic focus on fossil fuels, need for alternative energy sources, and financial and environmental implications of this transition.
- Attitudes and Reactions: This topic focus on frustration and anger about pollution and practices for better environmental awareness and responsibility.

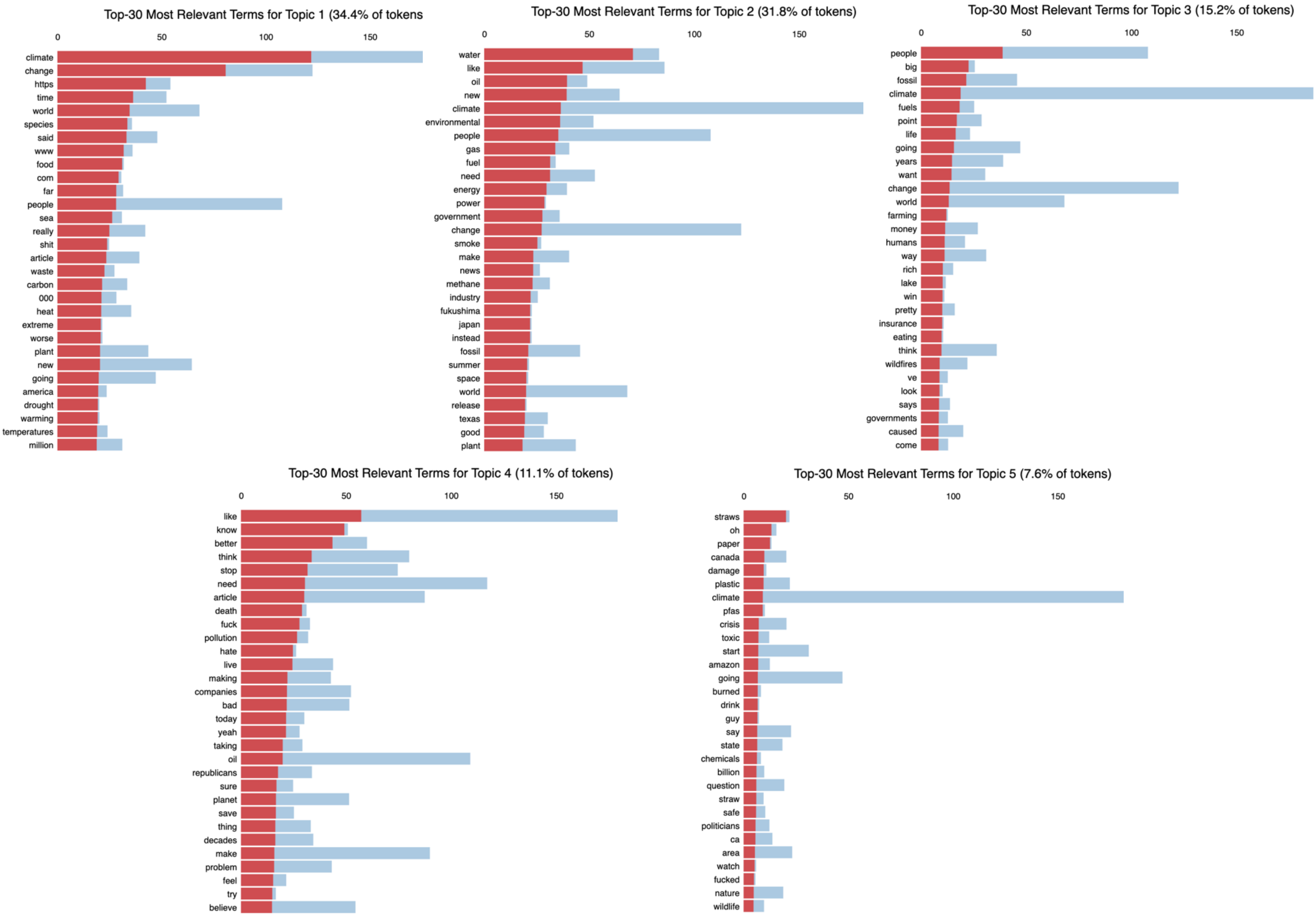


Figure 6 – Top 30 Most Relevant Terms for Each Topics

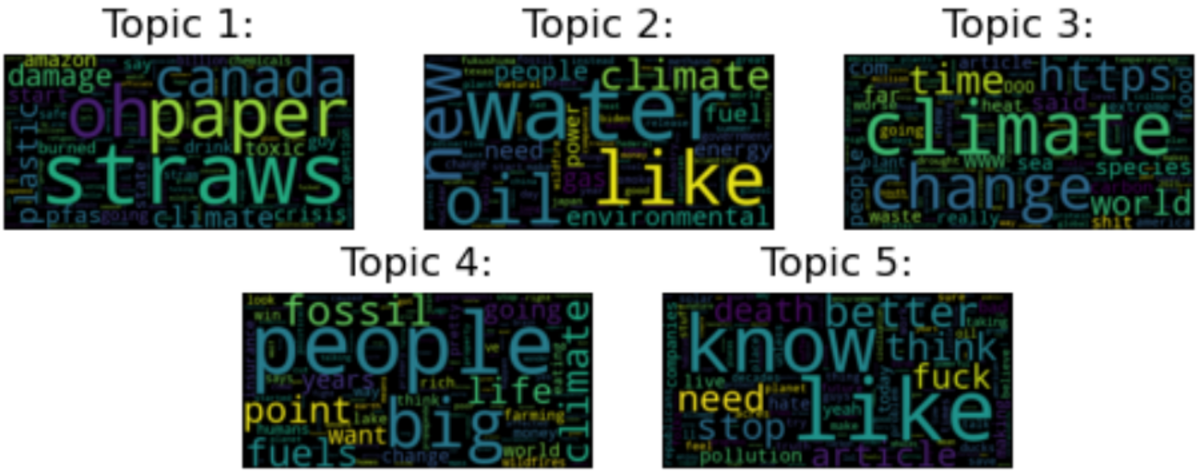


Figure 7 – Word Cloud of Each Topic

In Figure 6, the distribution of token percentages across each topic is presented with their relevant terms and Figure 7 shows their word clouds. Topic 5 exhibits the lowest token percentage at 7.6%. This percentage can be considered as appropriate for effectively identifying and understanding the content of the topic.

7 Conclusion

Through in-depth analysis of “environment”, it becomes evident that several trending topics have emerged among its community. Through LDA analysis, we define five distinct themes, covering areas such as climate change's impact on species, methane emissions, food production, space exploration, and climate data records. As well as climate change news, personal beliefs, rising sea levels, fuel sources, water-related issues, fossil fuels, sustainability, and events like Fukushima disaster. This subreddit serves as active platform for expressing feelings and opinions on environmental matters.

When considering the expression of feelings and opinions in posts and comments, the sentiment analysis conducted using the “Vader” method reveals a balanced emotional spectrum. The community’s responses to environmental matters encompass a mixture of positive, negative, and neutral sentiments. This shows that subreddit serves as platform for diverse emotional expressions, ranging from optimistic and hope to concern and critique. The analysis also shows that active and vibrant engagement of community in expressing their viewpoints, emotions, and opinions regarding various environmental topics.

8 References

1. Gupta, S. (n.d.). Text Data Pre-processing Techniques in ML. EnjoyAlgorithms. Retrieved from https://www.enjoyalgorithms.com/blog/text-data-pre-processing-techniques-in-ml
2. Kapadia, S. (2019, April 15). Topic Modelling in Python: Latent Dirichlet Allocation (LDA). Towards Data Science. Retrieved from https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0
3. Prabhakaran, S. (n.d.). Topic Modelling with Genism (Python). Machine Learning Plus. Retrieved from https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/
4. Chan, J. (2023). RMIT Canvas. COSC 2671: Social Media and Network Analytics. Retrieved from https://rmit.instructure.com/courses/107293/modules
5. Tran, S. (2020, November 7). Topic Modelling for Absolute Beginners. Retrieved from https://thedigitalskye.com/2020/11/07/topic-modelling-for-absolute-beginners/#:~:text=Looking%20at%20the%20chart%20on,to%20decreasing%20order%20of%20prevalence.