Sentiment Analysis of Reddit Data on Climate Change Compared to News Articles: A Comparative Study

Saketh Balaji Maddineni
Department of Computer Science
University of Exeter
Exeter, UK
Sm1304@exeter.ac.uk

Dr. Tristan Cann
Department of Computer Science
University of Exeter
Exeter, UK
T.J.B.Cann@exeter.ac.uk

Prof. Hywel Williams
Department of Computer Science
University of Exeter
Exeter, UK
H.T.P.Williams@exeter.ac.uk

***Abstract*—Climate change is a significant global issue widely covered by the media. Despite this extensive media coverage, the portrayal of climate change by the media may not always align with the general public's perceptions. The aim of this research, conducted by a single researcher, is to explore the differences in how climate change is perceived and discussed by the general public compared to the media by analyzing data from Reddit, a popular social news platform. Specifically, natural language processing (NLP) techniques, precisely sentiment analysis with VADER, will be utilized to analyze Reddit posts related to climate change during specific events in recent history, such as the 2016 and 2020 US presidential elections. Comparing the sentiment expressed in Reddit posts from subreddits actively used for discussions on climate change to that of news articles on climate, this research will offer insights into how these two media sources perceive climate change. This research will contribute to a better understanding of the public's attitudes towards climate change and the role of the media in shaping these attitudes. Through analysis, the aim is to identify potential areas for improving communication and awareness of climate change issues, ultimately leading to a more informed and engaged public.**

I. Introduction

Climate change is a pressing global issue that affects every living being, and the media have widely covered it. However, the media's portrayal of climate change may not always align with the general public's perceptions. This discrepancy can significantly impact how people understand and respond to climate change, making it a concern for researchers and policymakers alike.

The main objective of this study is to explore the differences in how climate change is perceived and discussed by the general public compared to the media by analyzing data from Reddit, a popular social platform and news articles related to climate change during significant events in the recent past.

Specifically, I will use sentiment analysis with VADER to analyze Reddit posts related to climate change during specific events in recent history, such as the 2016 and 2020 US presidential elections and Environment conferences COP2021 and COP 2022. We will also analyze news articles on climate during these events for comparison purposes.The subreddits used for the analysis are r/climate, r/climatechange, r/sustainability, and r/environment. The sentiment analysis will provide us with a measure of the emotional tone of the language used in the posts and news articles, enabling us to compare the attitudes towards climate change expressed in these two media sources. This project aims to contribute to a better understanding of the public's attitudes towards climate change and the role of the media in shaping these attitudes. Comparing the sentiment expressed in Reddit posts to news articles on climate during specific events, we will gain insights into how these two media sources perceive climate change. Moreover, I will compare the sentiment analysis of Reddit posts between similar events to showcase how people's perception of climate change has evolved. Ultimately, this research aims to identify potential areas for improving communication and awareness of climate change issues, leading to a more informed and engaged public.

II. Research Context

Climate change is one of the most pressing global challenges of our time. Public opinion and sentiment towards climate change are critical in shaping policy decisions and public awareness of the issue. Social media platforms like Reddit have emerged as important sources of public discourse on climate change, with millions of users discussing the topic daily. One area of research that has gained considerable interest in recent years is sentiment analysis of social media data. Social media platforms like Reddit, Twitter, and Facebook provide a valuable data source for analyzing public sentiment on climate change. As a result, sentiment analysis of social media data has become a popular research area, intending to understand public perception and sentiment toward climate change. Previous studies have used sentiment analysis to understand public attitudes toward climate change and track changes in sentiment over time. However, most of these studies have focused on a single social media platform or a specific event or period, like the articles mentioned below.

S.Kumar et al. [1] proposed a sentiment analysis decision system for tracking climate change opinions on Twitter. This study used machine learning techniques to classify tweets

on climate change as positive, negative, or neutral, providing insights into the sentiment of climate change discussions on social media. Similarly, B.Thapa et al. [2] conducted sentiment analysis of cybersecurity content on Twitter and Reddit. The study showed that both platforms had similar sentiment trends for cybersecurity topics, with Reddit having a higher degree of negativity compared to Twitter.

Regarding Reddit data extraction, S.Lindskog et al. [3] developed The Pushshift Reddit Dataset, a large, publicly available dataset of Reddit posts and comments that can be used for various research purposes. K.Kapur et al. [4] conducted a comparative study of sentiment analysis for multi-sourced social media platforms, including Reddit, and proposed a hybrid method that achieved better accuracy.

To gain insights into Reddit discussions on climate change, K.Treen et al. [5]analyzed the discussion of climate change on Reddit and investigated whether the discourse was polarized or deliberative. They found that while there were instances of polarization, there were also instances of productive deliberation. Additionally, B.Ruhoff [6] proposed a sentiment analysis model for Reddit, which was trained and evaluated on a dataset of Reddit comments.

Finally, to better understand attitudes towards climate change, Z Mi et al. [7] used sentiment analysis to analyze attitudes towards climate change on Twitter. They found that emotions and sentiments expressed in tweets about climate change are closely related and can be used to identify attitudes towards climate change.

Based on these studies, my research aims to contribute to a better understanding of the public's attitudes towards climate change and the role of the media in shaping these attitudes. By using sentiment analysis with VADER as said in [8] and data extraction from Reddit, I plan to explore differences in how climate change is perceived and discussed by the general public versus the media during specific events in recent history, such as the 2016 and 2020 US presidential elections and the Environment conferences COP2021 and COP2022. My study also compares sentiment analysis of Reddit data between similar events to showcase how people's perceptions regarding climate change have changed over time.

III. Aims Objectives

As mentioned earlier, the main idea of this study is to explore the differences in how climate change is perceived and discussed by the general public compared to the media by analyzing data from Reddit, a popular social platform and news articles related to climate change during significant events in the recent past. To achieve this idea, the following research questions and objectives have been identified:

A. Research Questions

1) How has public sentiment towards climate change on Reddit evolved during significant events, such as the 2016 and 2020 US presidential elections and the 2021 and 2022 COP conferences?
2) How does public sentiment towards climate change on Reddit compare to the sentiment of news articles published during the same period?
3) What are the implications of the findings for understanding public sentiment towards climate change and the role of news media in shaping this sentiment?
4) How have people's sentiments towards climate change transformed over time on reddit by comparing the sentiment during the similar events in mentioned in research question 1?

B. Objectives

1) Collect and preprocess Reddit data related to climate change from subreddits such as r/climatechange, r/climate, r/environment, r/climatescience, and r/climatepolicy during the events mentioned in research question 1.
2) Perform Perform sentiment analysis on the collected Reddit data and the news articles data using Natural Language Processing (NLP) techniques and tools such as VADER (Valence Aware Dictionary and Sentiment Reasoner) to determine the sentiment of the Reddit posts related to climate change during each significant event.
3) Analyze and visualize the trend of the sentiment of the Reddit posts towards climate change over time, using tools such as Matplotlib and Pandas.
4) Analyze and visualize the trend of the sentiment of the news articles towards climate change over time, using tools such as Matplotlib and Pandas.
5) Compare the sentiment of the Reddit posts with the sentiment of the news articles about climate change published during the same period, using tools such as Matplotlib and Pandas, and identify any discrepancies between the two.
6) Evaluate the effectiveness of news media in accurately capturing public sentiment towards climate change and discuss the implications of the findings for understanding public sentiment towards climate change and the role of news media in shaping this sentiment.
7) Identify any significant changes or patterns in public sentiment towards climate change on Reddit over time by comparing sentiment during similar events across different time periods.
8) Interpret the findings to gain insights into how the public's perception of climate change has transformed over time and provide recommendations for future research.

Overall, this project aims to contribute to our understanding of public sentiment towards climate change and the role of news media in shaping this sentiment by analyzing sentiment across multiple periods and events and comparing the sentiment of Reddit posts with news articles. The project will also explore how the sentiment of Reddit posts and news articles may differ during events such as the UK heatwave and the COVID-19 pandemic.

IV. Data Resources

For this project, the primary data sources will be Reddit posts and comments related to climate change. The posts will be collected using the Python Reddit API Wrapper (PRAW), which allows access to the Reddit API. The comments on the posts will be collected using the Pushshift API, which provides access to a historical archive of Reddit comments. The data collected will be limited to the following subreddits: climate change, environment, climate policy, climate action plan, sustainability, climate science, and climate. Additionally, the data will be limited to specific periods, including the 2016 and 2020 US presidential elections, the 2021 and 2022 COP conferences..The news artilces data will also be collected for the same time period from the news api. Regarding software tools, we will use Python as the primary programming language. We will also use the VADER sentiment analysis tool to perform sentiment analysis on the collected data. The project will be conducted on a local machine with adequate resources to handle the volume of data.

V. Methods Experiment Design

To achieve the stated objectives, the following methods will be used:

1) Data Collection: For Reddit data extraction, the Python Reddit API Wrapper (PRAW) will be used to obtain posts on climate change from subreddits r/climatechange, r/climate, r/environment, r/climatescience, r/climatepolicy, and r/sustainability during the 2016,2020 US presidential elections and 2021,2022 COP Conferences. Pushshift API will use the extracted post ID from PRAW to obtain comments related to posts. News articles about climate change published during the same events will be extracted from newsapi.ai.
2) Data Preprocessing: The collected Reddit data and news articles data will be cleaned by removing any unwanted characters, numbers, punctuations, and special symbols, other irrelevant words using natural language processing (NLP) techniques. For vectorization, the Term Frequency-Inverse Document Frequency (TF-IDF) or count vectorization method will be used to convert the text data into a matrix of numerical values that can be analyzed.
3) Sentiment Analysis:
   We will use a pre-trained sentiment analysis model to analyze the sentiment of both Reddit posts and news articles. Specifically, we will use the VADER (Valence Aware Dictionary and sEntiment Reasoner) model, a lexicon and rule-based sentiment analysis tool shown to perform well in social media sentiment analysis tasks. The VADER model assigns a score to each post/article, indicating the text's positivity, negativity, and neutrality.
4) Data Analysis:
   Once sentiment scores have been assigned to each post/article, we will analyze the data to answer our research questions and objectives. We will use Python libraries such as pandas and matplotlib to perform statistical analysis and visualization of the data. Specifically, we will compare the sentiment of Reddit posts across different periods and events with the sentiment of news articles published during the same periods. We will also evaluate the effectiveness of news media in accurately capturing public sentiment toward climate change.
5) Evaluation Plan:
   To evaluate the effectiveness of our sentiment analysis approach, we will randomly select a subset of posts/articles and manually annotate their sentiment. We will compare our sentiment analysis results with the manual annotations to calculate our approach's accuracy, precision, and recall. We will also evaluate the results qualitatively to ensure that they make sense from a human perspective.
   Overall, the methods described in this section will be used to collect, preprocess, and analyze data to answer our research questions and objectives.

VI. Data Governance Ethics

Data collection and analysis will be done in accordance with ethical standards, and all necessary precautions will be taken to ensure the privacy and security of the information. The project will adhere to the guidelines established by the Institutional Review Board and the General Data Protection Regulation. Data will be collected from Reddit using the PRAW and PUSHSHIFT API tools. These tools will collect data based on keyword inputs and will only access publicly available data. No personal data will be collected or used in the analysis. The collected data will be stored in a secure location with access granted only to authorized personnel. The data will be anonymized and aggregated to prevent the identification of individual users. [9] As for the ethical concerns, the study will be conducted with respect for the principles of non-maleficence, beneficence, respect for autonomy, and justice. The study will not cause any harm to the participants or subjects, and their privacy and confidentiality will be respected. The study will also avoid any form of bias or discrimination. [9] In conclusion, this section outlines the project's data governance and management plan and the ethical considerations to be considered. The project will comply with all legal and ethical guidelines, ensuring the privacy and security of all information involved.

VII. Project Plan

The following is a project plan that outlines the milestones and deliverables of the project, along with their due dates:

1) Week 1 (Apr 20 - Apr 26): Literature review and research proposal
2) Week 2-3 (Apr 27 - May 10): Data collection
3) Week 4-5 (May 11 - May 24): Data preprocessing and cleaning
4) Week 6-7 (May 25 - Jun 7): Development of sentiment analysis model

5) Week 8-9 (Jun 8 - Jun 21): Sentiment analysis of Reddit data
6) Week 10-11 (Jun 22 - Jul 5): Comparison of sentiment analysis results with news articles
7) Week 12 (Jul 6 - Jul 12): Analysis of sentiment trends during different periods
8) Week 13-16 (Jul 13 - Aug 10): Report writing and submission

The project will be monitored by regular meetings with the project supervisor, who will provide guidance and feedback. Any delays or issues that arise during the project will be addressed and resolved promptly to ensure the project is completed on time. Assigned enough backup time in the above stated project plan to make sure everything in the pipeline goes on smoothly.The project will be closely monitored to ensure its successful execution.

VIII. Risk Assessment

As with any project, there are potential risks that could prevent the successful completion of the project. The following is a risk assessment for this project, along with possible mitigation strategies:

1) Data Availability: There is a risk that the required data may not be available or may be limited in scope, which could hinder the project's progress. Multiple data sources such as different subreddits as backup to those mentioned earlier will be identified to mitigate this risk, and efforts will be made to obtain data from each source.
2) Data Quality: There is a risk that the data obtained may be noisy or incomplete, which could negatively impact the sentiment analysis. To mitigate this risk, data preprocessing and cleaning techniques will be applied to ensure quality data.
3) Sentiment Analysis Accuracy: There is a risk that the developed sentiment analysis model may need to accurately capture the sentiment of the Reddit data, leading to incorrect conclusions. To mitigate this risk, the model will be evaluated using appropriate metrics, and adjustments will be made if necessary.
4) Ethical Issues: There is a risk of ethical issues arising, such as privacy violations, partial data, or unintended consequences of the sentiment analysis. A data governance plan will be developed to mitigate this risk, and ethical considerations will be considered throughout the project. [9]
5) Technical Issues: There is a risk of technical issues, such as software or hardware failures, which could delay or disrupt the project. Backup plans will be developed to mitigate this risk, and regular backups will be made to prevent data loss.

In conclusion, above are the expected risks and possible proposed mitigations. Nevertheless, there will always be a chance for unexpected issues during the project. To handle these issues, enough time has been planned as a buffer at every level of the project plan.

IX. Conclusion

In conclusion, this proposed study aims to explore the differences in how the general public perceives and discusses climate change compared to the media. By analyzing data from Reddit posts and news articles about climate change during significant events in recent history, sentiment analysis will be used to compare the emotional tone of language used in these two media sources. This project seeks to contribute to a better understanding of the public's attitudes towards climate change and the role of the media in shaping these attitudes. Moreover, by comparing sentiment analysis of Reddit posts between similar events, we will showcase how people's perception of climate change has evolved. Ultimately, the findings of this research may help identify potential areas for improving communication and awareness of climate change issues, leading to a more informed and engaged public. This project holds the potential to provide valuable insights into the complex interplay between media and public perception, ultimately contributing towards the global efforts to mitigate the impact of climate change.

References

[1] S. Kumar, N. A. Jailani, A. Singh, and S. Panchal, "Sentiment analysis on online reviews using machine learning and nltk," *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 1183–1189, 2022.

[2] B. Thapa, "Sentiment analysis of cybersecurity content on twitter and reddit," *ArXiv*, vol. abs/2204.12267, 2022.

[3] S. Lindskog and J. A. Serur, "Reddit sentiment analysis," *DecisionSciRN: Stock Market Decision-Making (Sub-Topic)*, 2020.

[4] K. Kapur and R. Harikrishnan, "Comparative study of sentiment analysis for multi-sourced social media platforms," *ArXiv*, vol. abs/2212.04688, 2022.

[5] K. M. d'I. Treen, H. T. P. Williams, S. J. O'Neill, and T. G. Coan, "Discussion of climate change on reddit: Polarized discourse or deliberative debate?" *Environmental Communication*, vol. 16, pp. 680 – 698, 2022.

[6] B. Ruhoff, "Sentiment analysis for reddit," *Technical Library*, vol. 254, 2016. [Online]. Available: https://scholarworks.gvsu.edu/cistechlib/254

[7] Z. Mi and H. Zhan, "Text mining attitudes towards climate change: Emotion and sentiment analysis of the twitter corpus," *Weather, Climate, and Society*, 2023.

[8] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI: AAAI*, pp. 216–225, 2014. [Online]. Available: https://www.aclweb.org/anthology/W14-21.pdfpage=234

[9] OpenAI, "Chatgpt: An ai language model for natural language processing," *OpenAI Blog*, 2021. [Online]. Available: https://openai.com/blog/chat-gpt-3/

[10] B. Dahal, S. A. Kumar, and Z. Li, "Topic modeling and sentiment analysis of global climate change tweets," *Social Network Analysis and Mining*, vol. 9, pp. 1–20, 2019.

[11] J. Yao, "Automated sentiment analysis of text data with nltk," *Journal of Physics: Conference Series*, vol. 1187, 2019.

[12] A. A.A., A. M.O., and O. O.L., "Sentiment analysis of reddit comments," *International Journal of Computer Science and Mobile Computing*, 2022.

[13] M. Lydiri, Y. E. Habouz, and H. Zougagh, "Sentiment analysis decision system for tracking climate change opinion in twitter," in *Conference on Business Informatics*, 2022.

[14] S. Li, Z. Xie, D. K. W. Chiu, and K. K. W. Ho, "Sentiment analysis and topic modeling regarding online classes on the reddit platform: Educators versus learners," *Applied Sciences*, 2023.

[15] L. Barros, A. Trifan, and J. L. Oliveira, "Vader meets bert: sentiment analysis for early detection of signs of self-harm through social mining," in *Conference and Labs of the Evaluation Forum*, 2021.