Mentor Name. : John Rachlin
Subject/Category : Data Science & Sustainability
Class : DS3500, Session A - Track 3
Group # : 12

Perspectives on a Sustainable Future

Priyanka Adhikari, Ruchira Banerjee, Nidhi Bendre, Simone Kaplunov

Background

Our project has a focus on analyzing YouTube comments from TED Talks related to sustainability, utilizing public discussions as a resource to gain more insights. The project helps to uncover how sustainability is framed and perceived within public discourse, recognizing that these conversations can significantly influence awareness, policy, and individual actions towards sustainable living. By employing advanced text analysis techniques across various visualizations—including sentiment analysis, word frequency comparisons, and more—we aim to offer a comprehensive snapshot of the current sustainability dialogue.

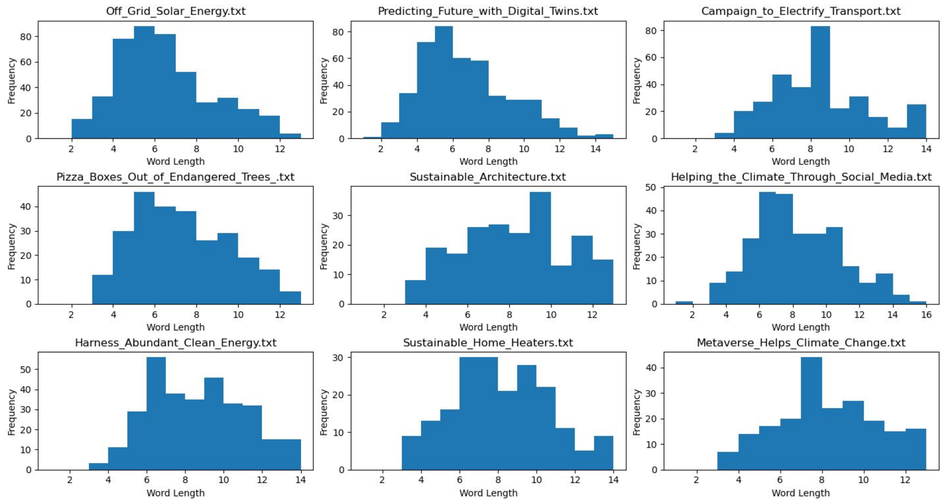
Methods

We began the project by consolidating comments from YouTube videos into individual text files. Each of the files were parsed through to eliminate punctuation as well as common stop words. This allowed only relevant text to be selected, and the words were standardized to be all lowercase. We used various Python packages and libraries including plotly, string, and collections, alongside our existing sankey code to collect comprehensive statistics from the text data. We collected key metrics such as word count, word length distributions, and sentiment scores. From there developed specific functions to visualize our data, such as histograms to highlight word length distribution, a bar chart for sentiment analysis, and sankey diagram to illustrate common words across the different texts.

Results

Fig. 1: Word Length Distribution

Most comments show a preference for shorter words, which indicates a broader accessibility of the language used in discussions. Certain topics have a wider distribution of word lengths, which points to more technical vocabulary being employed by the commenters.



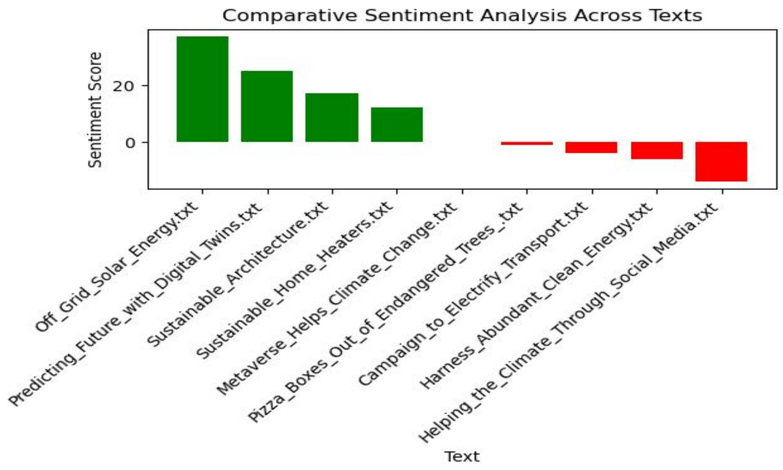
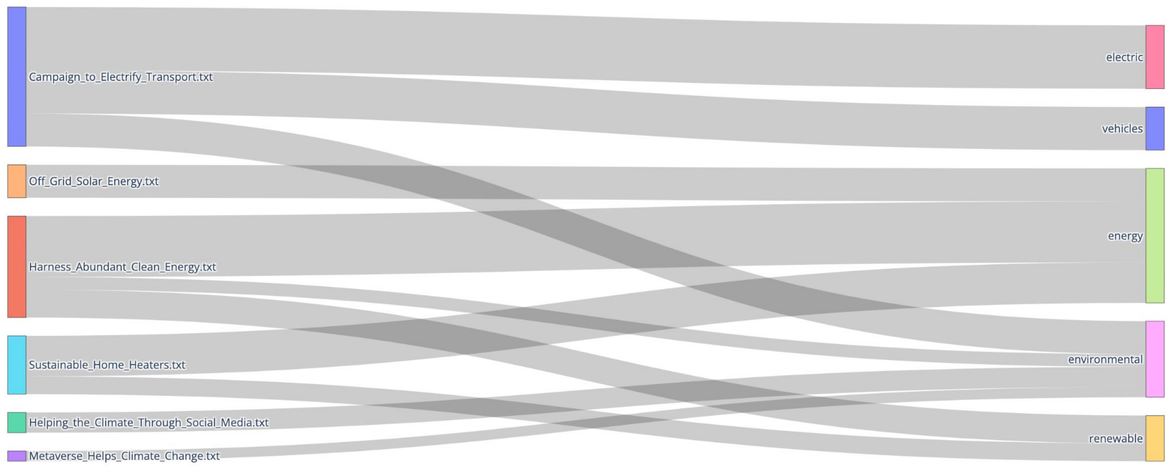


Fig. 2: Sentiment Analysis

High variability in sentiments highlights the complexity of audience reactions to different sustainability themes. Positive remarks on certain TED Talks can be a reflection of relatability of the topic and the presenter's delivery.

Fig. 3: Sankey Diagram

The Sankey diagram illustrates common themes in sustainability discussions, with wider strands representing more frequent mentions of terms like "electric" and "energy". This suggests high public engagement with these topics. The interlinking of terms indicates the interconnected nature of sustainability in public conversations.



Conclusion

Most comments have a preference for shorter words due to the online, informal nature of the commentary. There was high variability in sentiment for the videos, indicating the importance of the video subtopic and speaker's abilities. The most common words were *energy*, *environmental*, *electric*, and *solar*, so these terms should be actively used in new videos to engage viewers.

Team Contributions

Priyanka: sentiment chart, Conclusion
Ruchira: sankey visualization, Methods
Nidhi: text loading & stop words, Results
Simone: word length histogram, Background