NASA Climate Change Sentiment Analysis project

SECTION 1: INTRODUCTION & PROJECT OVERVIEW

1.1 Project Background

Climate change remains one of the most polarized topics in global discourse.[1] As a leading scientific authority, NASA's public communications serve as a focal point for both scientific inquiry and public skepticism.[2] This project analyzes the digital footprint of climate change discussions on NASA's social media platforms to bridge the gap between scientific communication and public perception.

1.2 Problem Statement

Large-scale public engagement often results in high-volume, unstructured text data. For organizations like NASA, manually monitoring sentiment or identifying key areas of public concern is impossible. There is a critical need for an automated system to:

1. Quantify the emotional tone of the public.
2. Identify the specific topics that generate the most interaction.
3. Predict engagement levels to optimize future communication strategies.

1.3 Objectives

- **Sentiment Quantification:** Classify public comments into Positive, Neutral, or Negative categories.
- **Topic Discovery:** Use Latent Dirichlet Allocation (LDA) to extract the "hidden" themes within thousands of comments.
- **Predictive Modeling:** Develop a Machine Learning model to forecast LikesCount based on textual features.
- **Strategic Insights:** Provide data-driven recommendations for improving public science engagement.

SECTION 2: DATA ARCHITECTURE & METHODOLOGY

2.1 Data Collection & Description

The dataset comprises over 500 high-engagement comments from NASA's Facebook page spanning 2020–2023.[3]

- **Key Features:** Date, LikesCount, CommentsCount, and Text.
- **Data Integrity:** User privacy was maintained via SHA-256 hashing of profile names.

2.2 Data Preprocessing (ETL Pipeline)

To prepare the unstructured text for machine learning, the following pipeline was implemented:

1. **Noise Reduction:** Removal of URLs, special characters, and non-ASCII symbols.

2. **Tokenization & Normalization:** Converting text to lowercase and breaking it into individual words (tokens).
3. **Stop-Word Filtering:** Removing common words (e.g., "the", "is") using the English stop-word dictionary to focus on "content words" like "CO2", "atmosphere", and "warming".[4]
4. **Vectorization:** Utilizing CountVectorizer to convert text into a numerical matrix for model ingestion.[5]

2.3 Analytical Framework

- **Sentiment Analysis:** Applied **TextBlob** to calculate Polarity scores ranging from -1 (Extremely Negative) to +1 (Extremely Positive).[6]
- **Machine Learning:** Implemented a **Random Forest Regressor** to model the non-linear relationship between comment structure and user engagement (LikesCount).[7]

SECTION 3: ANALYSIS & RESULTS

3.1 Sentiment Analysis Results

The analysis revealed a predominantly **Positive to Neutral** sentiment across the dataset.

- **Positive Sentiments:** Often associated with "Thank you" messages, curiosity about NASA's imagery, and appreciation for "visualizing the invisible."
- **Negative Sentiments:** Primarily focused on climate skepticism or frustration regarding the perceived slow pace of global policy changes.

3.2 Topic Modeling (LDA) Insights

The LDA model successfully clustered the discourse into three primary themes:

1. **The Atmospheric Theme:** Discussions revolving around CO2 levels, the stratosphere, and greenhouse gases.
2. **The Skepticism/Debate Theme:** Comments questioning historical temperature graphs (e.g., the last 10,000 years) or dismissing data as "fake."
3. **The Educational Theme:** Appreciation for NASA’s ability to explain complex scientific phenomena to the general public.

3.3 Model Performance

The **Random Forest Regressor** (stored as climate_engagement_model.pkl) showed that textual length and specific keywords (e.g., "CO2", "science") are strong predictors of engagement. The model effectively identified that comments providing "neat comparisons" or asking deep scientific questions often receive higher "Like" counts than short, generic comments.

SECTION 4: STRATEGIC RECOMMENDATIONS & CONCLUSION

4.1 Strategic Recommendations

Based on the data, the following strategies are recommended for corporate/organizational communication:

- **Encourage Visual Storytelling:** Comments praising "visualizing the invisible" suggest that data visualization is NASA's strongest engagement tool.
- **Address Specific Queries:** High engagement was found in comments asking about the "stratosphere" and "troposphere." Creating dedicated "FAQ" posts on these specific layers could boost engagement.
- **Moderate Constructively:** Since the skepticism theme is prominent, automated sentiment monitoring can help identify and address "denialist" trends early with fact-based responses.

4.2 Conclusion

This internship project successfully applied advanced Data Science techniques to a real-world communication challenge. By integrating NLP and Machine Learning, we transformed 500+ raw comments into a roadmap for public engagement. The ability to predict comment popularity and categorize themes allows for a more proactive approach to scientific communication in the digital age.
