SENTIMENT ANALYSIS OF FACTORS AFFECTING CLIMATE CHANGE

Abinaya Mohan
Student ID: 33085799
Module: Applied Artificial Intelligence
Sheffield Hallam University, Department of Computing
April 22, 2024

1 Abstract

We kept an eye on what people were saying about climate change on social media, particularly Twitter. Why? Because combating this worldwide issue requires a grasp of public opinion. We concentrated on tweets addressing greenhouse gases, global warming, and carbon emissions. We de-contaminated the data, examined the sentiments expressed in the tweets, and developed a model to quantify those sentiments. This can influence policy and increase public understanding of climate change.

2 Introduction

Climate change is a huge issue, and understanding what people think about it is key. This project listened in on social media, especially Twitter, to see what people are saying about carbon emissions, greenhouse gases, and global warming. We cleaned up the data and analyzed the emotions behind the tweets to get a sense of public opinion. This can help raise awareness and influence policies aimed at tackling climate change.

3 Methodology

We dove into social media, particularly Twitter, to see what people are saying about climate change. We snagged tweets mentioning carbon emissions, greenhouse gases, and global warming. To really understand the conversation, we cleaned up the data and prepped it for analysis. Finally, we used different tools to analyze the sentiment – the overall feeling – of the tweets. By comparing these tools, we aimed to find the best way to grasp the emotions behind what people are saying.

3.1 Data Collection and Knowledge Extraction

We collected tweets about climate change by focusing on hashtags and keywords related to carbon emissions, greenhouse gases, and global warming. Then, we dug into these tweets using various tools to understand the conversation: how people felt (sentiment analysis), what words were most common (word clouds), what topics were trending, how sentiment changed over time, and how different factors related to each other.

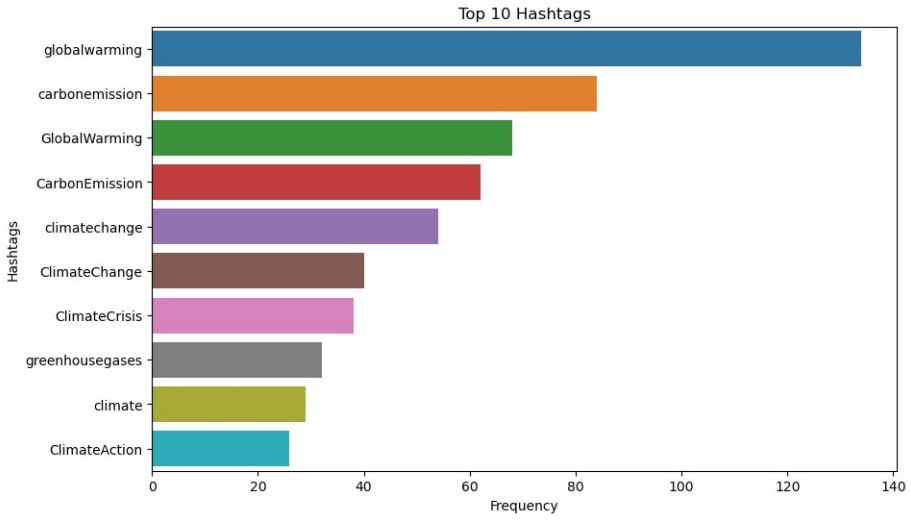

Figure 1: Most popular hashtags about climate change

	Name	Handle	Retweets	Likes	Tweet URL	Profile Link	Post Body	Timestamp	searchQuery
0	Ernest Markham	@ErnMarkham	0.0	0.0	https://twitter.com/ErnMarkham/status/17256786...	https://twitter.com//ErnMarkham	GLOBAL WARMING is waiting for no one. Surely w...	2023-11-18T00:53:00.000Z	global warming
1	Jonathan Overpeck	@GreatLakesPeck	37.0	83.0	https://twitter.com/GreatLakesPeck/status/1725...	https://twitter.com//GreatLakesPeck	Yet more confirmation that climate change is h...	2023-11-17T12:04:00.000Z	global warming
2	∞ Vishy ∞	@VishalManve12	1.0	0.0	https://twitter.com/VishalManve12/status/17132...	https://twitter.com//VishalManve12	"There is increasing evidence that global warm...	2023-10-14T16:35:00.000Z	global warming
3	Good Law Project	@GoodLawProject	1182.0	1735.0	https://twitter.com/GoodLawProject/status/1725...	https://twitter.com//GoodLawProject	NEW INVESTIGATION\n\nWe can reveal that, just...	2023-11-17T14:23:00.000Z	global warming
4	Prem Sikka	@premnsikka	131.0	173.0	https://twitter.com/premnsikka/status/17256237...	https://twitter.com//premnsikka	Environment secretary Steve Barclay received d...	2023-11-17T21:15:00.000Z	global warming

Figure 2: Overview of Collected tweets

3.1.1 Distribution of Sentiment

This research on emotions in climate change tweets gives a snapshot of how people are feeling. By looking at a score that combines different factors, we can classify tweets as positive or negative. This picture gives a clear idea of the overall sentiment and opens the door for a deeper dive.

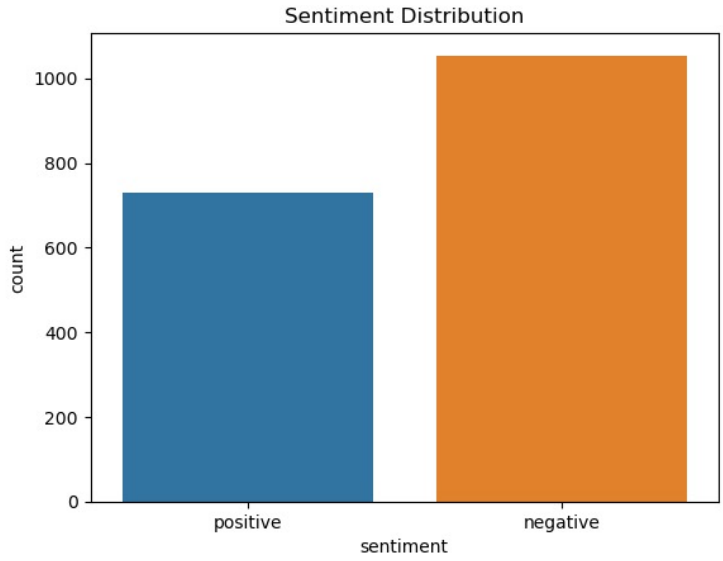

Figure 3: Sentiment

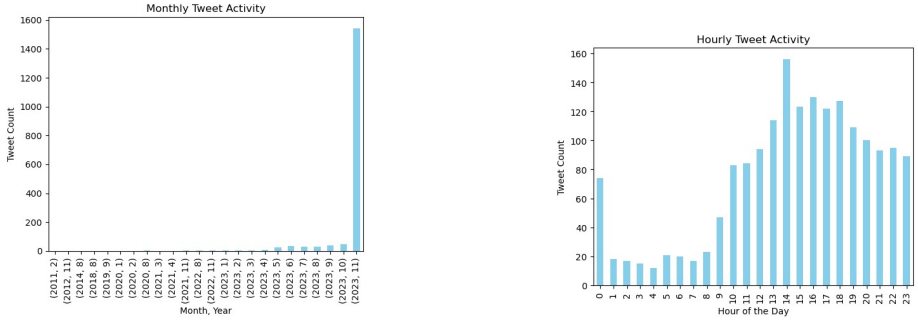

Figure 4: Public Interest Fluctuation over time

3.1.2 Word clouds

Word clouds are a great way to see what people are talking about most when it comes to climate change. We can analyze positive and negative tweets to compare the most frequent words. This helps us understand the key arguments and terms used by both sides of the issue.

Figure 5

3.2 Text Normalization Methodologies

Text normalization is the process of standardizing the format and minimizing noise in text data to prepare it for analysis *(Topic modeling and sentiment analysis of global climate change tweets, 2019)*. Two essential methods are employed in the context of sentiment analysis on Twitter data: stemming and tokenization.

3.2.1 Tokenization

Tokenization is a technique we utilized to extract the individual words from each tweet. It was similar to dissecting every word in every tweet. Understanding and analyzing the tweets' content is a crucial first step.

- Managing Punctuation:

  To improve our analysis of social media sentiment, we removed punctuation that didn't add meaning to the overall feeling of the text. This helps us focus on the words that matter most when gauging emotions.

- Implementation:

  A special function was written to sort through the punctuation marks in text data. This function figured out which punctuation helped understand the overall feeling of the text, and kept those marks. The NLTK library provided a helpful set of tools to identify these punctuation marks in the English language.

- Eliminating Stopwords:

  To get a clearer picture of emotions in a social media text, we removed common words that don't affect the overall feeling. This helps the analysis focus on the meaningful words that express sentiment.

- Alphanumeric Filtering:

  To improve the sentiment analysis of Twitter data, we removed extra noise from the text. We built a tool to focus only on important letters and numbers within each tweet. This helps the analysis target the most relevant parts of the conversation.

3.2.2 Stemming

Text cleaning isn't complete without a step called stemming! This process boils words down to their most basic form, like their root. By doing this, similar words (like "happy," "happiest," and "happiness") get grouped. This helps streamline the analysis by reducing the number of features the computer model needs to consider, ultimately improving efficiency *(Lexicon-Based sentiment analysis and emotion classification of climate change related tweets, 2022)*.

- Implementation: To get the best results from sentiment analysis, we explored different stemming techniques. NLTK's stemmer modules, PorterStemmer, LancasterStemmer, and SnowballStemmer, were all put to the test. We compared the sentiment analysis results after applying each technique to see which one led to the most accurate understanding of the overall feeling in the text data.

- Methods of Stemming:

  For text analysis, NLTK provides three different stemmers: PorterStemmer, LancasterStemmer, and SnowballStemmer. PorterStemmer takes a more conservative approach, applying a set of principles to eliminate suffixes and uncover a word's underlying form. In contrast, LancasterStemmer is more forceful in breaking down words into their most basic form by employing a distinct set of criteria. Last but not least, SnowballStemmer addresses stemming from a multilingual standpoint and provides a well-rounded solution based on a reliable algorithm. We assessed each of the three methods to determine how they affected the sentiment analysis findings.

- Assessing Under- and Over-stemming:

  We tested different ways to shorten words (stemming) to see which worked best for analyzing emotions in climate change tweets. Stemming can be tricky - you don't want to lose too much meaning. We tried three methods and compared them to see which gave the most accurate understanding of the overall feeling in the tweets.

4 Model Training and Evaluation for Sentiment Analysis

During this stage, logistic regression—a popular approach for binary classification problems like sentiment analysis—was utilized to build sentiment analysis models. Before the use of logistic regression, the text data was subjected to diverse text vectorization techniques to convert them into a format that is appropriate for machine learning methods.

Text vectorization, which transforms textual input into numerical vectors that machine learning algorithms can understand, is an essential stage in sentiment analysis. Three primary text vectorization techniques were used in this

project: Positive/Negative Frequency, Count Vectorization, and Term Frequency-Inverse Document Frequency (TF-IDF).

4.1 Count Vectorization

This technique counts the instances of every word in a document to represent text data. It generates a sparse matrix in which every row denotes a document and every column a distinct word across the corpus. The frequency of each word in the corresponding documents is shown by the values in the matrix. This method just takes into account a word's appearance in the document, ignoring its order. Count Vectorization is renowned for its ease of use and effectiveness in capturing the corpus's vocabulary *(Barachi, 2021)*.

- Accuracy Calculation

  To evaluate the model's performance on the Count Vectorized test data, we used an accuracy score function. This approach is similar to how Positive/Negative Frequencies are used for sentiment analysis. Essentially, the function compares the model's predictions with the actual sentiment labels in the test data and calculates a score based on how often it's correct.

- Confusion Matrix

  Using a plotting function, we were able to generate a visual depiction of the confusion matrix to gauge the model's performance. The model's predictions are broken out in detail in this visualization, which also displays the proportion of tweets that were correctly and mistakenly labeled as positive, negative, etc.

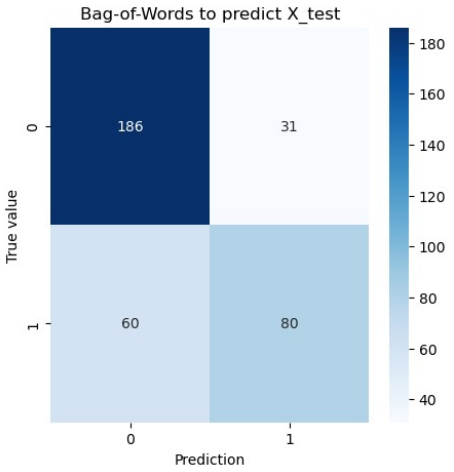

Figure 6: LR Model Accuracy : 74.15%

4.2 TF-IDF (Term Frequency-Inverse Document Frequency)

The statistical metric known as TF-IDF is used to assess a word's significance in a document about a corpus. In addition to considering a word's frequency within a document (term frequency), it also considers the word's rarity throughout the entire corpus (inverse document frequency). Because they are thought to be more informative, words that are common in a document but uncommon in the corpus are given higher weights. By minimizing the influence of common words that appear often in all publications, TF-IDF seeks to capture the distinctiveness of terms in a given document.

- Accuracy Calculation:

  We used the accuracy score function once more to assess the model's performance on the TF-IDF vectorized test data. The way this function operates is by contrasting the sentiment labels included in the test data with the sentiment predictions generated by the model. Based on how frequently the model's predictions matched the true sentiment, it assigns a score.

- Confusion Matrix

  The model's accuracy on TF-IDF data was assessed with an accuracy score and visualized using a confusion matrix for better understanding.

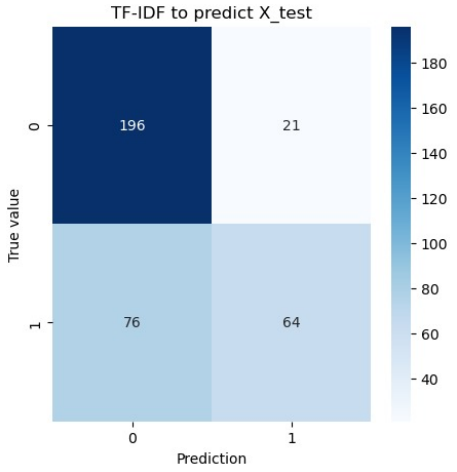

Figure 7: LR Model Accuracy : 72.83%

4.3 Positive/Negative Frequency

This method counts the instances of words that convey both positive and negative sentiment on each page. It entails creating lists of terms that are both

positive and negative and figuring out how frequently they appear in the text. This method gives a general idea of the sentiment of the document by quantifying the presence of sentiment-bearing terms.

- Accuracy Calculation

  To measure how well the model performed, we compared its sentiment predictions for the test data with the actual sentiment labels. This gave us a numerical score representing the model's overall accuracy.

- Confusion Matrix

  A chart called a confusion matrix, helped us see the model's strengths and weaknesses on climate change posts. It showed where the model excelled at predicting sentiment and where it made mistakes.

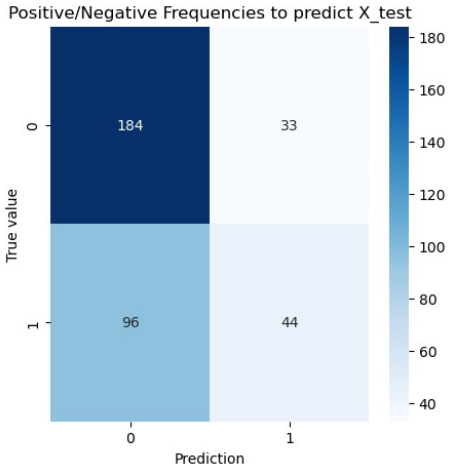

Figure 8: LR Model Accuracy : 63.87%

5 Results and Discussion of Sentiment Analysis Models on Batch Data

Analyzing climate change posts revealed Count Vectorization works best to capture the sentiment. This highlights the importance of text preparation for accurate social media analysis.

5.1 Model Performance Variation

Testing different ways to prepare social media text for sentiment analysis showed Count Vectorization captured emotions best in climate change discussions.

5.2 Simplicity vs. Complexity

Count Vectorization, a simpler method that focuses on word counts, surprisingly did the best at capturing sentiment in climate change posts. This highlights the interesting trade-off between complex models and task-specific effectiveness.

5.3 Contextual Challenges

The Positive/Negative Frequency method wasn't as accurate, highlighting the challenge of understanding sentiment in social media. Discussions about climate change often use layered language and ever-evolving ideas. Capturing the full range of emotions behind these posts might require more intricate tools.

Analyzing climate change posts on social media taught us a key lesson: how we prepare the text data hugely impacts results. We'll use this to improve our methods and explore stronger tools to capture the emotions behind these complex conversations.

References

Bryan-Smith, L., Godsall, J., George, F., Egode, K., Dethlefs, N., & Parsons, D. R. (2023). Real-time social media sentiment analysis for rapid impact assessment of floods. *Computers & Geosciences*, 178, 105405. https://doi.org/10.1016/j.cageo.2023.105405

Barachi, M. E., Alkhatib, M., Mathew, S. S., & Oroumchian, F. (2021). A novel sentiment analysis framework for monitoring the evolving public opinion in real-time: Case study on climate change. *Journal of Cleaner Production*, 312, 127820. https://doi.org/10.1016/j.jclepro.2021.127820

Dahal, B., Kumar, S., & Li, Z. (2019). Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining*, 9(1). https://doi.org/10.1007/s13278-019-0568-8

Fagbola, T. M., Abayomi, A., Mutanga, M. B., & Jugoo, V. (2022). Lexicon-Based sentiment analysis and emotion classification of climate change related tweets. In Lecture notes in networks and systems (pp. 637–646). https://doi.org/10.1007/978-3-030-96302-6_60

Li, W., Haunert, J., Knechtel, J., Zhu, J., Zhu, Q., & Dehbi, Y. (2023). Social media insights on public perception and sentiment during and after disasters: The European floods in 2021 as a case study. *Transactions in GIS*, 27(6), 1766–1793. https://doi.org/10.1111/tgis.13097