Mini Project 3: Sentiment Analysis on Twitter Climate Change Data

Mudita Shakya

1. Introduction

Climate change is a growing concern all over the world. It is important to take steps now so that eventual consequences such as: decrease in the quality of life, increase in diseases, and mass migrations can be prevented. Although the scientific evidence points to climate change and global warming being an urgent problem, there is still a significant chunk of the human population that believes that climate change is a hoax. Nowadays, a significant amount of data is generated from discussions carried out on social media. Twitter is one such source of textual data that can be used to monitor the trends and opinions of people.

Monitoring the content and number of tweets on climate change over time can help us understand the concerns of the people and which topics are mostly discussed. Keeping track of twitter discussions can help us identify any spikes in negative sentiment or feelings of non-belief towards climate change which in turn can lead us to taking more steps to generate awareness. To perform this, we can use sentiment analysis techniques to analyze and track people's opinions on climate change.

In this project, two approaches: supervised learning with different implementations and classification with a pretrained model were used to predict the sentiment expressed in tweets collected over a period of 10 years.

2. Data Processing

The following steps were carried out for performing data processing:

2.1. Collecting the data

The data used for the project was collected using the Twitter API. The tweets were collected for a period of 10 years starting from 1st January 2013 to 15th October 2022. Since historical data was required, I applied for and received the Academic research access. This allowed me to use the search-all API [a] which is only approved for those with academic research access. To collect the tweets related to climate change, the keywords: climate change and global warming were used along with the hashtags: #climatechange and #globalwarming for the search. The following was used as the query for the search:

"climate change OR global warming OR #climatechange OR #globalwarming lang:en"

For each year, around 5000 tweets were collected resulting in a total of around 50,000 tweets. Different details about the tweets were collected such as: author id, creation time, geo location id, tweet id, language, like count, reply count and retweet count along with the actual tweet text.

2.2. Labeled Dataset

For training the supervised models, we also require a labeled dataset. For this purpose, I used the Twitter Climate Change Sentiment Dataset from Kaggle [b]. This dataset contains 43943 tweets related to climate change collected between April 27, 2015, and February 21, 2018. The labels are of the following classes:

i. **News (2)**: tweet that is linked to factual news about climate change
ii. **Pro (1):** tweet that supports the belief of man-made climate change

iii. **Neutral (0):** tweet that neither supports nor refutes the belief of man-made climate change

iv. **Anti (-1):** tweet that does not believe in man-made climate change

The data distribution in the labeled dataset was as follows:

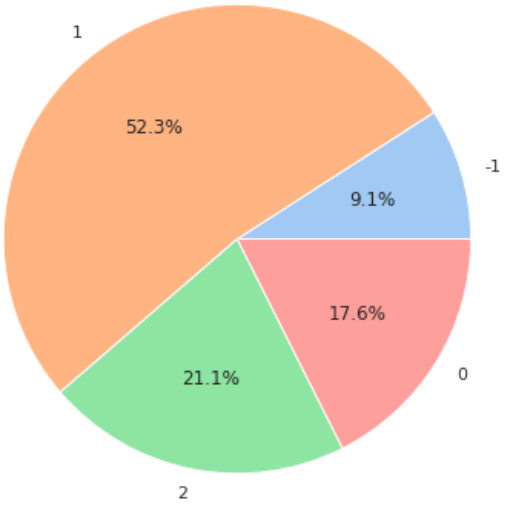

Fig. 1: Pie chart of the data distribution of the sentiment labels

Sentiment	Count
1	22962
2	9276
0	7715
-1	3990

Table 1: Table showing the count of the different labels in the labeled dataset

2.3. Data Preprocessing

Before using the data for model training and prediction purposes, we need to preprocess the data. The dataset was checked for null values and no null values were found. It was important to remove duplicate tweets as well since a lot of the tweets might just be retweets. After removing duplicates, a total of 26905 unique samples were remaining.

For the preprocessing, the following steps were taken:

i. Changing the tweet text to lowercase

ii. Removing mentions from the tweet

iii. Removing hashtags from the tweet

iv. Removing URLs from the tweet

v. Removing retweet indication from the tweet

vi. Removing special characters from the tweet

The step to remove stop words was left out on purpose because in general, stop words also contain words like not, hasn't, didn't, and so on which could contribute to the sentiment expressed in the text. The cleaned tweet was added as a new column called "cleaned_tweet" and another column called "year" was added to distinguish the year in which the tweet was published. The data preprocessing steps were applied in both the labeled dataset and the extracted tweets dataset.

3. Model Training

The cleaned_tweet column is then transformed to a matrix of TF-IDF features using TfidfVectorizer from scikit-learn. Then the data from the labeled dataset is split into train and test datasets using scikit-learn with a ratio of 75:25 for the train and test respectively, resulting with the following counts:

Dataset	Total count
Train Dataset	32957
Test Dataset	10986

Table 2: Table showing the train and test split

Different approaches were tested out to find the best one suited for the task. The following supervised learning techniques were used and compared:

i. Random Forest Classifier

ii. K Nearest Neighbor (kNN) Classifier

iii. Logistic Regression

iv. Linear Support Vector Machine Classifier

v. Extra Trees Classifier

Also, a pretrained model was tested to check and see how well it performs on this dataset. The pretrained model used was the **VADER** sentiment analysis tool. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media [c]. Since the VADER implementation gives out only 3 labels: positive, negative, and neutral, the labels: News and Neutral were combined. The VADER tool did not perform well with the data, resulting in an accuracy of only 35.28%.

Model	Accuracy
Random Forest	0.6862
kNN Classifier	0.4584
Logistic Regression	0.6565
Linear Support Vector Machine (SVM) Classifier	0.6838
VADER (pretrained)	0.3528
Extra Trees Classifier	0.7021

Table 3: Table showing the accuracy comparisons between different trained and one pretrained model

3.1. Findings

- The Extra Trees Classifier was found to give the best performance with 70.21% accuracy.
- The Extra Trees Classifier also outperformed the pre-trained model (VADER) tool. This could also be because the labels for this particular use case are a little different from the normal sentiment analysis use case.
    - Normally, sentiment analysis tasks have the labels: positive, negative, and neutral.
    - In this case, the labels are assigned based on whether the tweet author believes in man-made climate change.

4. Prediction

Since the Extra Trees Classifier was found to give the best performance among the tested models, the trained Extra Trees Classifier was used to perform the classification for the extracted tweets data. The following was the distribution of the sentiments in the prediction results.

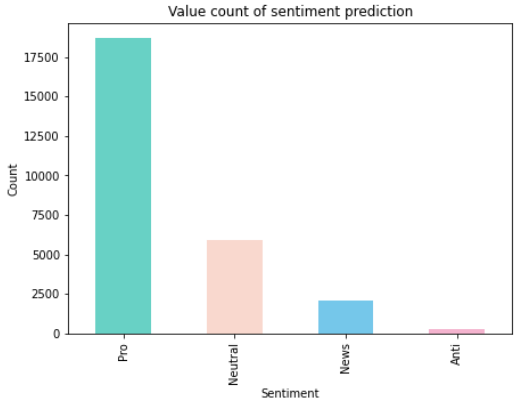

Fig. 2: Bar chart showing the distribution of the predictions across the different sentiment labels

Prediction	Count
1	18701
0	5895
2	2058
-1	251

Table 4: Table showing the count of the different sentiment labels in the prediction for the extracted dataset

As seen in the figures, most of the tweets (69.5%) were predicted to be Pro and 0.9% of the tweets were predicted to be Anti. The rest of the tweets were classified to be neutral and those conveying news.

The distribution of sentiments in tweets by year is shown as follows:

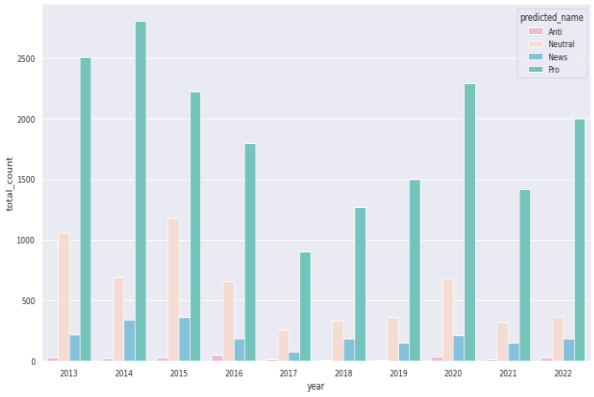

Fig. 3: Bar graph showing the sentiment polarity over the 10-year period

From this, we can observe that most tweets are from those who believe in man-made climate change. We can see the number of tweets expressing non-belief in climate change is low but has spikes in 2016, 2020, and 2022.

I also used the **geopandas** library to convert the location in the tweet data to latitude and longitude values. To identify where the significant chunk of the people who deny climate change come from, the latitude and longitude values of the tweets that were predicted as being "Anti" were plotted on the world map using a scatter plot.

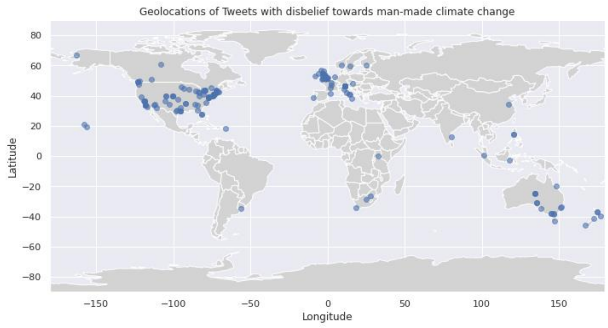

Fig. 4: Scatter plot on the world map showing the geo locations of the tweets labeled as "Anti"

The plot shows that the negative tweets are dense around North America (specifically United States) and north-western Europe. These areas can be focused on to generate more awareness regarding climate change.

5. Further Analysis

I performed some analysis to further my understanding of the nature of the discussions being carried out on the topic of Climate change on Twitter.

5.1. Identifying key words

To analyze the data further, the top 20 words used in the tweets data was found. This was done by first preparing a list with all the word of the tweet. The words were then stemmed using SnowballStemmer from NLTK. The stemming process helps to get the root of the word. Then according to the number of occurrences in the tweets, the top 20 words were plotted using a bar graph.

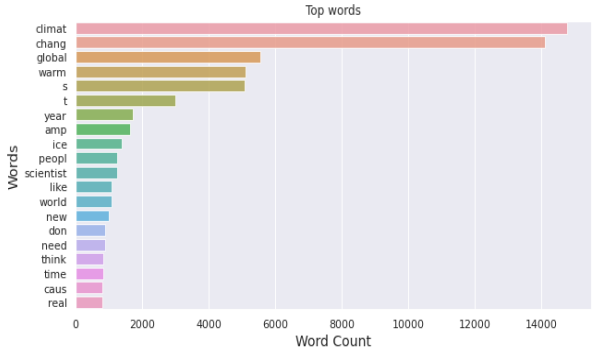

Fig. 5: Bar plot showing the word count of the top 20 words used in the tweets

The top 15 hashtags were found for each of the sentiment labels to further understand the nature of the tweets in the extracted data.

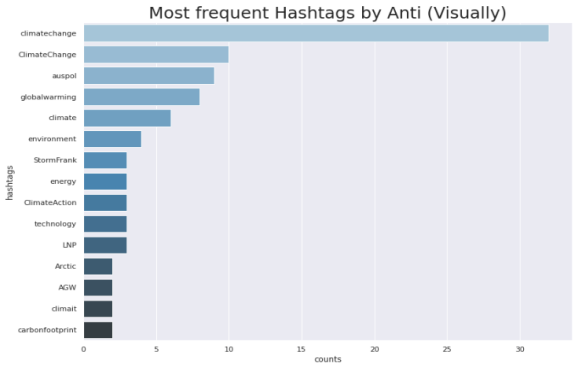

Fig. 6: Bar plot showing the count of the top 15 hashtags found for the "Anti" sentiment label

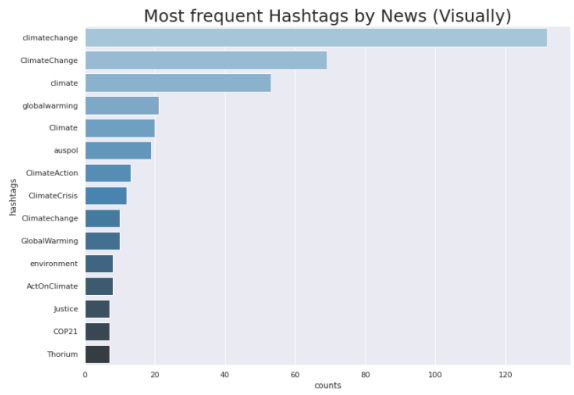

Fig. 7: Bar plot showing the count of the top 15 hashtags found for the "News" sentiment label

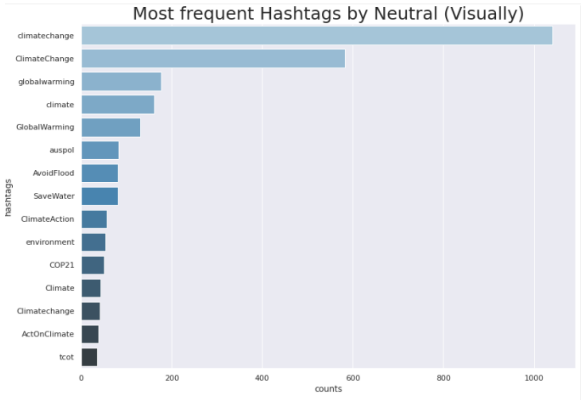

Fig. 8: Bar plot showing the count of the top 15 hashtags found for the "Neutral" sentiment label

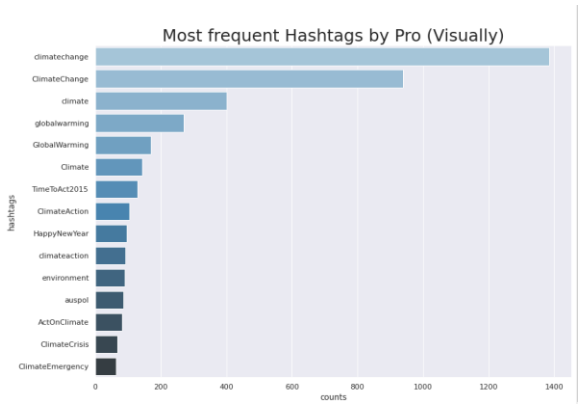

Fig. 9: Bar plot showing the count of the top 15 hashtags found for the "Pro" sentiment label

5.2. Topic Modeling with LDA

To further understand the data and the context of the discussions carried out on the topic of Climate Change, we can use topic modelling techniques such as Latent Dirichlet Allocation (LDA) to cluster the discussed text into different topics. Topics are collection of prominent keywords which help to describe what the topic is about. For performing topic modeling, the tweets were first tokenized. The collection of words was then converted to a bag of words and converted to a corpus. The LDAMulticore model from the genism library was used and trained on the prepared corpus and used to model 15 topics from the data.

The results from this were visualized using the pyLDAvis library. The pyLDAvis library helps to create an interactive visualization of the results from the LDA model. Each bubble in the visualization represents a topic. The most important terms identified for the topic is also displayed.

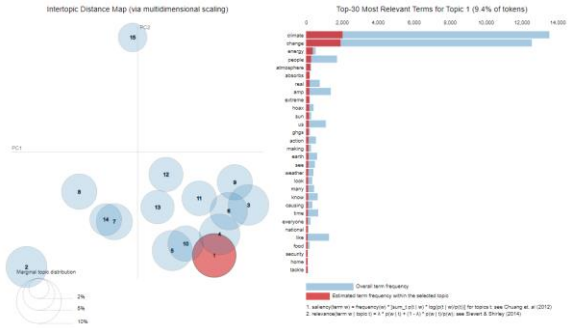

Fig. 10: Visualization showing the topics modeled by LDA

This helps to identify and learn about the main topics being discussed regarding climate change and provides an insight into the discussions being carried out on the general topic of climate change.

5.3. Twitter Retweet Network

To understand the network created by these discussions, I used NetworkX to visualize the retweet network of the tweets. Only the tweets

with a retweet count greater than 500 were considered for the network. Each tweet author was taken as a node and each account mentioned in the tweet was also considered as a node and edges connected the retweet author to the original author of the text. This resulted in a network with 673 nodes and a network graph as seen below:

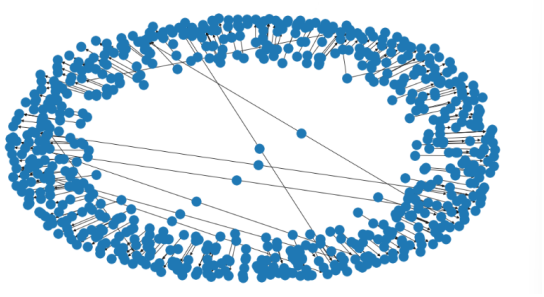

Fig. 11: NetworkX visualization of the retweet network

6. Conclusion

There were some bottlenecks experienced in the collection of tweets for the project.

i. I experimented collecting tweets using the Tweepy library but I got varying number of tweets every time I ran the query, which is why I resorted to using the Twitter API directly. I also experimented with different query strings for getting the tweets. I was able to get good results with using both the keywords and the hashtags for the search.

ii. Another issue was that the labeled dataset that I used was not balanced. More than 50% of the labels were for the Pro sentiment. This affected the performance of the models that were trained on this data. Also, the labels used in this dataset are different from the standard labels used for sentiment analysis, which are positive, negative, and neutral. Therefore, for comparing the performance of the VADER tool, I had to combine the data for Neutral and News labels.

From the model training and evaluation, it was found that the Extra Trees Classifier provided the best performance out of all the models that were tested. The predictions from the trained Extra Trees classifier on the extracted tweets dataset revealed that most of the tweets expressed support for the belief of man-made climate change. By plotting the locations of the tweets predicted as "Anti" by the model, it was clear that most of such tweets came from the United States and north-western Europe. The hashtag "**climatechange**" was the most popular hashtag in tweets in all sentiment categories. This might have been influenced by the search query as well. From Topic Modeling, the words: climate, change, global, and warming were the most salient terms in all topics found.

7. References:

a. Twitter API Documentation: Twiiter API

b. Twitter Climate Change Sentiment Dataset: Kaggle Dataset

c. Hutto, C.J. & Gilbert, E.E. (2014). 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, June 2014.