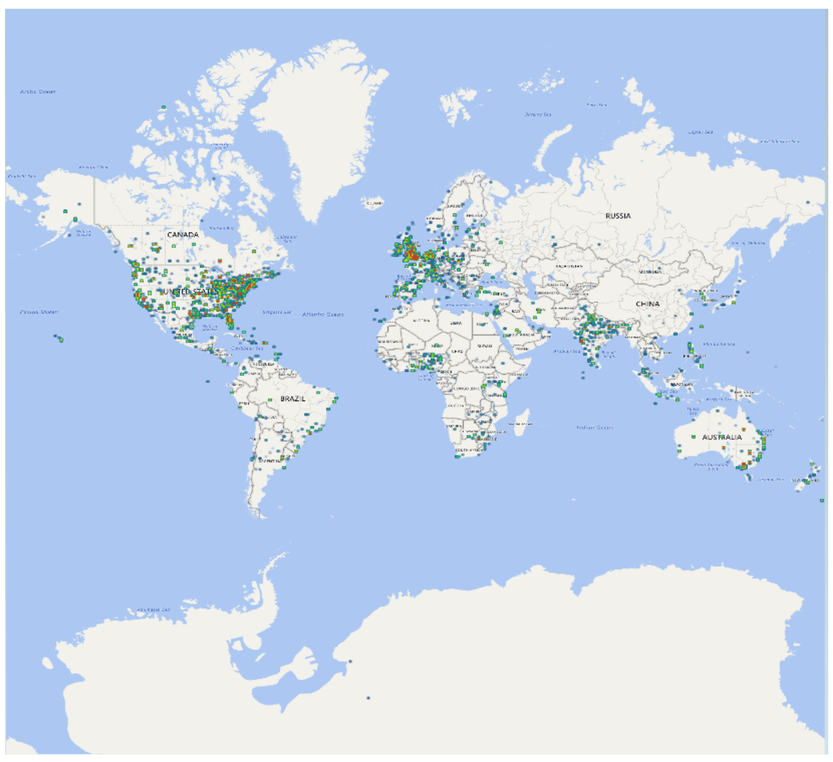
# Machine learning approach to large scale social media data analysis


## Abstract

This project intends to collect large data sets consisting of social media posts to analyze them using machine learning models. The analysis determines the public opinion on a given topic, namely environmental policy over time and in different places. During the initial stage of the project, a web-crawler has been implemented, and has performed data collection; then, we cleaned and pre-processed the collected data and assembled a reliable data set for further analysis; after that, the machine learning models was trained and tested with the collected data, to detect emotions, opinions, and more characteristics of social media posts.


## Twitter Data

The collection of data we managed to gather consists of abut 213000 Tweets posted by users between November, 30th 2018 and December, 3rd 2018. The bulk of the data was collected by using a Twitter Streamer with a Python application we developed. We then build the indexing and searching application with Python Package, Pandas to extract twitter messages that are written in English and that also mention climate change. This method has been instrumental in collecting a highly significant set of tweets, but it is only a random sub-sample of all the opinion-related content which are written in English and are related to climate change. Twitter's platform allows users to re-tweet the tweets that they believe might be interesting, this includes articles, news, and opinions. Note that our analysis is based on this climate Twitter collection including re-tweeted tweets. The reason for including re-tweets is because we assume that it detects the sentiment of users by re-tweeting tweets of other users. The heat map below represents the consistency of tweets across the world.

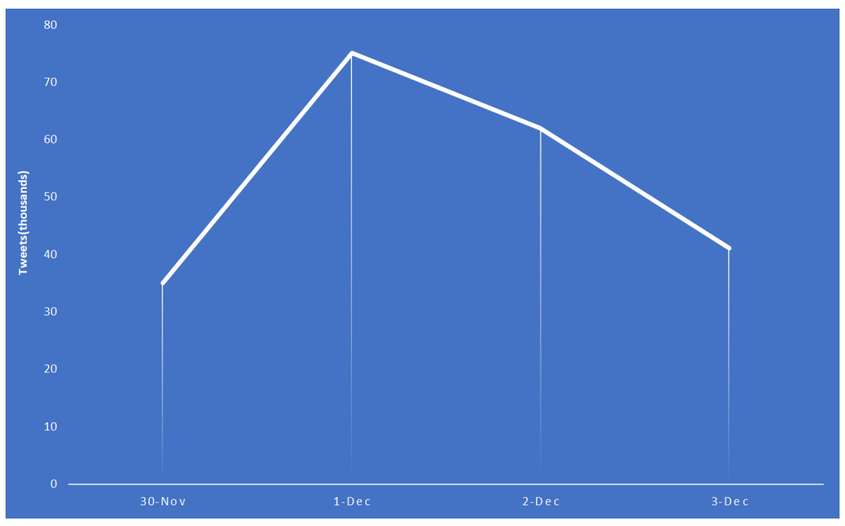



## Sentiment Analysis

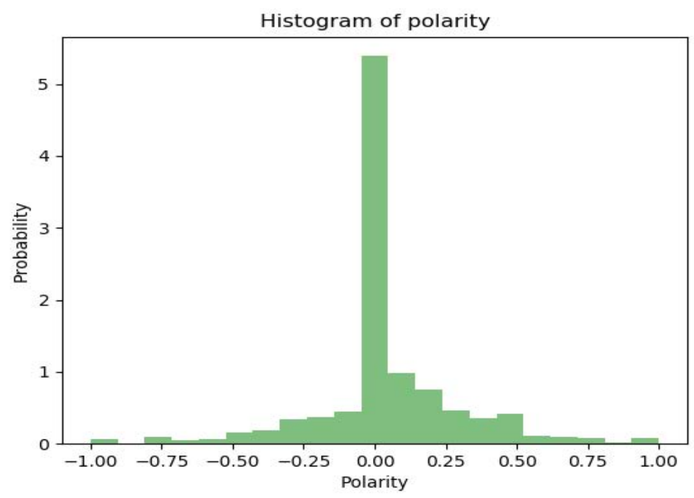
In this research, We explored a customized method for sentiment text classification using NLP to characterize the sentiment content of a text unit. We pre-processed our data as follows:

- We lowercased all letters (strip casing of all words)
- Tokenized (convert the string to a list of tokens based on whitespace and remove punctuation marks)
- Removed rare words ( suggests that words occurring two or fewer times may be removed, since these words are unlikely to be present to aid in future classifications)

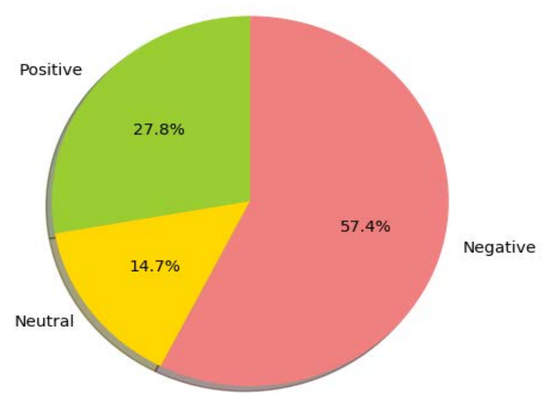
## Results and Event Detection




The intention of this extraction is to try to explore the percentage of climate change related tweets per day. There are a total of 213; 000 tweets related to climate change in our collection, with about 20; 000 climate change tweets daily on average. A plot of the percentage of tweets regarding climate change recorded daily is displayed in the Figure above. We observe that the percentages show high variability, and major fluctuations are also detected. For example, on December 1st, the percentage goes up extremely high but then goes down significantly within the next two day. The peak could be possibly explained by the occurrence of 2018 G20 Buenos Aires summit.




The data set percentages below present a normal distribution. This variability is influenced by many factors, such as the news, articles published on that special day or the occurrence of any event. Because of these confounding factors on the figure below, it is easy to detect major changes or event using the polarity. It would be quite beneficial to climate sentiment studies if we can detect whether the sudden change in Twitter sentiment regarding climate change are related to major climate events or extreme weather conditions. We, thus, focus on the sentiment polarity probability.




## Conclusion

Traditionally, the attitudes, knowledge, and opinions of citizens and key decision-makers have been studied through relatively expensive and logistically challenging survey techniques, but more recently scientists and many other groups have begun to exploit the vast amounts of information available in social media platforms. This paper presents proof of concept results to suggest that mining social media data, exemplified here through Twitter accounts, can be a valuable way to yield insights on climate change opinions and societal response to extreme events. Considering the variation in sentiment polarity shows that there is still significant uncertainty in overall sentiment. We used Twitter data to illustrate how the opinions of Twitter users can change over time and in the aftermath of specific events, but similar approaches may be extended to other publicly available information and social-media platforms.


HAMAD EL KAHZA, DAVID FERNANDO GAMBOA SALAZAR

Advisor: DR. HOUWEI CAO