Aisha B., Allen S., Ritika S.

CS301-102

Group 6

Professor: Dr. Pantelis, TA: Yashwee Kothari

Due Date: April 24, 2022

**Group Six Project:**

Partly Sunny With a Chance of Hashtag

**Websites**

<https://github.com/aishabhamla/dataScience>

<https://www.kaggle.com/competitions/crowdflower-weather-twitter/overview>

**Abstract**

As the years go by, social media platforms, such as Facebook, Instagram, and even Twitter, are being used more frequently. We are at a point where we can use these platforms to predict the weather and when the weather has occurred. In this project, our team was provided a data set with tweets from Kaggle, which was related to the weather. Using those tweets, we had to come up with a simple Python script that will not only predict what the weather is but will also predict when it happened (past, current, or future). In this project report, the reader will understand the problem we chose to work on, the approach we took to solve it, and our results. They will read about the different related works group six had looked into, the data that had been collected, methods that have been used, and how the experiment went, and also what the whole team learned and concluded at the end of this project.

**Introduction**

In the era of big data, social media applies many information methods. In this project, we will be provided a data set with tweets related to the weather. With these provided tweets, our group will create a simple Python script that will read these tweets and analyze them to determine the kind of weather, the time, and the sentiment. This project is to create a software that will analyze a set of tweets which will be given to the software itself and from there it will determine if the weather is one of the three sentiments (positive, neutral, or negative) and then tell if the referenced weather in the tweet occurred in the past, present, and or future. The main objective of this project is to be able to assign certain labels to tweets based on the weather information that we can learn from them. We chose to solve this problem using the classification approach. Using solutions like decision trees were taken into consideration because they fit our requirements. While the competition’s performance metric is based on how many labels are accurately assigned, we will be focusing on getting the outputs right.

**Related Work**

Many companies have done similar work towards this concept of predicting the weather and some have researched or are still researching with the use of social media. Space Weather Prediction Center Space (SWPCS) has Weather Scales that were introduced to communicate to the public about the current and future weather conditions. They analyzed different readings and also developed a computer model to predict the weather and its warnings better than the technology from before. The Space Weather Prediction Center Space (SWPCS) has assembled forecasts for different types of space weather phenomena to help predict more promising and reliable ones.

Weather is one of the biggest influences in our everyday lives. Regardless, as the concept of data analytics continues to expand, the role of social media in the future of weather prediction is becoming more notable. In a recent research completed by the Warwick University (located in Britain) found that certain words on social media can be used to predict weather conditions before they actually occur. Nataliya Tkachenko of the Warwick Institute had said “our analysis demonstrates that metadata in social media image postings enables them to be used as social sensors, which can serve as a valuable supplement to instrument-based systems for predicting and monitoring floods, and other kinds of natural hazards”. So other than the Space Weather Prediction Center Space (SWPCS), many institutes have done similar research that was very similar to ours.

**Data**

For this report, we will be working with numerous data sets (tweets) that are from Kaggle. The set of tweets consists of words such as “hurricane”, “hot”, “cold” “rainy”, etc. It consists of four files, and a total of 57 columns. The training set contains tweets, locations, and a confidence score for each of 24 possible labels. The 24 labels come from three categories: sentiment, when and kind. For example, if the labels s4, w4, k4, k5 refers to “Positive”, “past weather”, “hot”, “humid”, respectively, a tweet “The hot and humid weather yesterday was awesome!” could have s4=1, w4=1, k4=1, k5=1, with the rest of the labels marked as zero. The four files given are “sampleSubmission.csv”, “test.csv”, “train.csv”, and “variableNames.txt”. These files are uploaded on the GitHub link. Each of these contains important information which is why when we download it we use:

*kaggle competitions download -c crowdflower-weather-twitter*

We will be using many functions that we have used throughout the program that we have learned this semester that will be mentioned in the next section (the Methods Section).

**Methods**

Since this challenge is mainly dealt with keywords and phrases, we had to use certain methods. To start off, we imported many packages such as NumPy, Pandas, Matplotlib, and also JAX. Grid search was also used to search for the best matching parameters. The fit regression model is also needed to organize the details. Although NumPy and JAX go hand in hand, we believed that both functions should be used to make the project more readable, run faster, and would be easier to debug whenever there is an issue with compiling the program when it finishes.

The biggest challenge we encountered during this experiment was attempting to download the data and coding the Python program to analyze the data line by line. After we figured out how to solve that particular problem, the next obstacle that came our way was trying to have the program decide what its output should be on its own.

Some functions we utilized in the program include a sanitization function. This function in particular was important to make analyzing the tweets easier. Each tweet was stripped of any punctuation, links, or Twitter mentions and all the tweets were converted to lowercase. We used the NLTK library to remove stopwords like “@”, “@”, “link”, “google”, “Facebook”, “yahoo”, “rt”, etc.

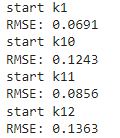
**Experiments**

For this experiment, we chose to use a fit regression model to determine the weather referenced in the tweets’ weather and whether the weather occurred in the past, present, or the future.

Part of The Given Data:



The Expected Output Would Be:



For reference, k1, k10, k11, and k12 refer to “clouds”, “rain”, “snow”, and “storms”, respectively. A storm might refer to heavy snow, floods, tornadoes, etc. Given the neutral tweet, the program was then able to associate it with labels that it might apply to, outputting the above labels and the mean squared error calculated for each label. For k1, the Root Mean Squared Error (RMSE) is 0.0691. This means that the program analyzed the tweet “There Was a Storm” and determined that it referred to a cloudy weather. A zero (0.0000) score would suggest that the tweet did not refer to cloudy weather. we were 100% in determining whether if there were clouds or not.

**Conclusion**

In conclusion, this project applied pattern and logistic recognition. In this competition, we were provided a set of tweets that were related to the weather. Data repositories sometimes have more in common with a landfill than a library. In this project, we learned that with python libraries such as NumPy, Pandas, Matplotlib, and also JAX, we can create a python program that could read a set of those tweets and predict the weather reading those tweets, and can tell when has/will it happen. After doing this project, we strongly feel that we are heading towards a new era where we can predict much more than just weather with the use of social media.

**References**

* NOAA US Department of Commerce. (n.d.). National Weather Service. Retrieved April 2022, from https://www.weather.gov/
* The Weather Company: The internet of weather things. Social Media for Business Performance. (n.d.). Retrieved April 2022, from https://smbp.uwaterloo.ca/2017/03/the-weather-company-the-internet-of-weather-things/