PROJECT REPORT

On

Sentiment and Weather Correlation

ISE 244 – AI Tools and Practice for Systems Engineering

Dr. Shilpa Gupta

shilpa.gupta@sjsu.edu

BY

Ketan Malempati

SJSU ID:

016695354

ketan.malempati@sjsu.edu

Master's in Artificial Intelligence

Table of Contents

1. Problem Definition

The impact of weather on human emotions and behavior is a widely researched area, and the use of social media data to analyze people's sentiments towards the weather has become increasingly popular in recent years. In this project, I propose to investigate the correlation between weather and sentiment within social media data. The objective of the project is to understand how weather conditions affect the sentiments expressed on social media. The current research uses machine learning models like SVM, BERT, rule-based, and naive-bayes methods for sentimental analysis and poses the following challenges:

- Handling multiple languages: Sentiment analysis across multiple languages poses unique challenges due to the variations in grammar, syntax, and cultural differences.
- Irony and sarcasm: Irony and sarcasm can be difficult to detect in text, as they often involve the opposite meaning of what is actually being said.
- Ambiguity and subjectivity: Sentiment analysis can be difficult because language is often ambiguous and subjective. Words can have multiple meanings and can be interpreted differently depending on the context and the person reading them.

1.1. Introduction

The weather's impact on human emotions and behavior has been extensively researched, and in recent years, there has been an increase in interest in examining how people feel about the weather using data from social media.

Weather is an integral part of our daily lives and affects us in numerous ways. Our emotions and behavior are also closely tied to the weather conditions we experience. It is a well-known fact that weather can affect our daily routines, decision-making, and overall well-being. For instance,

a sunny day may make us feel more positive and energized, while a gloomy day may make us feel sad and lethargic.

The weather has been shown in studies to have a variety of effects on our mood. Understanding how weather affects our emotions and behavior can assist us in developing strategies to improve our well-being, productivity, and, ultimately, quality of life.

Nowadays, Twitter, Facebook, and Instagram are used a lot. Social media platforms are a rich source of information about people's sentiments and behaviors as people share their thoughts and comments without any restrictions or hesitation, making them an ideal tool to study the impact of weather on human emotions.

This project examines weather data and social media sentiment to see if there is any relationship between them. We will analyze tweets and their effects on individual moods and behaviors.

I would be fetching Twitter data from Twitter using their API and also fetching the weather of the place at the time the tweet originated. I would then use the roBERTa-base model, a transformer-based model trained on a dataset of approximately 58 million tweets and fine-tuned for sentiment analysis. I will then be using decision trees, random forests, or ensemble methods to determine the correlation between sentiment and weather in a particular location. The goal of this project is to use machine learning models to look into the relationship between weather and sentiment in social media data. By handling multiple languages, irony and sarcasm, ambiguity, and subjectivity in sentiment analysis, as well as using the roBERTa-base model, the project hopes to find out how weather affects how people feel on social media. The proposed method, which includes analyzing the data with decision trees, random forests, or ensemble methods, appears to be an effective way to reach the project objective.

Finally, the goal of this research is to improve our understanding of the complex relationship between mood and weather, as well as to inform strategies for promoting positive emotional health in various weather conditions.

1.2. Literature Review

In the article "Mood and Weather: Feeling the Heat?" they have compared the sentiment of different states in the US and their correlation with the weather. In another paper "Sentiment Analysis of Twitter Data for Predicting Stock Market Movements," they used the sentiment to determine whether the market would rise or fall, which is very similar to our approach. In this article, we find the correlation and then predict the weather based on the sentiment.

The approach proposed in "Speaking of the Weather: Detection of meteorological influences on Sentiment within social media" has been applied in various contexts, including the study of tourists' perceptions of climate. One article that cited the selected article is "Tourists' Perceptions of Climate: Application of Machine Learning to Climate and Weather Data from Chinese Social Media" by Y. G. Tao, F. Zhang, W. J. Liu, AND C. Y. Shi published in 2021.

In the study, they use sentiment analysis and topic modelling to identify tourists' attitudes and opinions towards weather-related risks based on social media data from Sina Weibo. The authors note that the approach proposed in "Speaking of the Weather" is limited by potential biases in social media data and the need for further research to improve the accuracy of sentiment analysis.

2. Project Objectives

By fulfilling these goals, the project hopes to find the correlation between weather and sentiment. So that we can understand how weather affects our emotions and behaviour which could assist us in developing strategies to improve our well-being, productivity, and, ultimately, quality of life.

- Follow the process of experiments conducted in the paper "Speaking of the Weather: Detection of meteorological influences on Sentiment within social media" by JS. Zimmerman and U. Kruschwitz on a dataset created by me and also different methodologies.
- Evaluate the performance of various models like decision tree, XGBOOST, Random

Forest, Neural Networks, and Ensemble method.

- Find the correlation between sentiment and weather conditions.
- Explore ways to improve the accuracy of the model.
- Conduct a comparative analysis of the performance of different models and identify the strengths and weaknesses of each approach.
- Ensure the model is efficient and scalable to data that can be used in real-world applications.
- Develop a detailed documentation of the project including the data preprocessing steps, model architecture, hyperparameters used, and experimental results.
- Ensure the project follows ethical considerations such as privacy, fairness, and transparency.

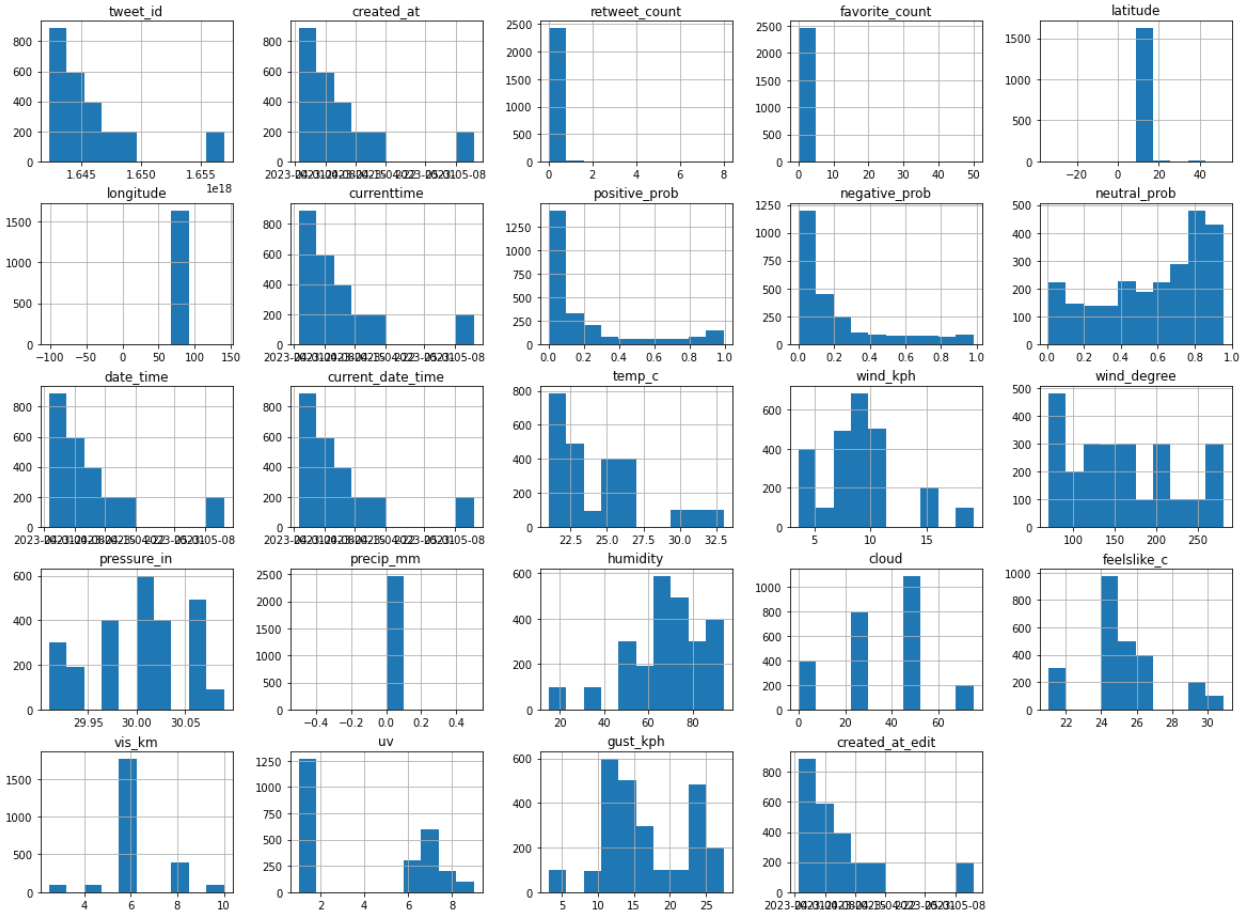
3. Data

Twitter Data - Tweets were collected over the period using the Twitter API. A filter was put in to get only tweets within the United States and the English language. We get features like ID, tweets, favourites, language, location, retweets, and many more. A filter was sent in the request to return only geo-tagged tweets in a bounding box that included tweets from San Jose and Bangalore. Additional filtering was included to remove any tweets, not within those cities and any languages that were not English. Only tweets where the coordinates were provided were used.

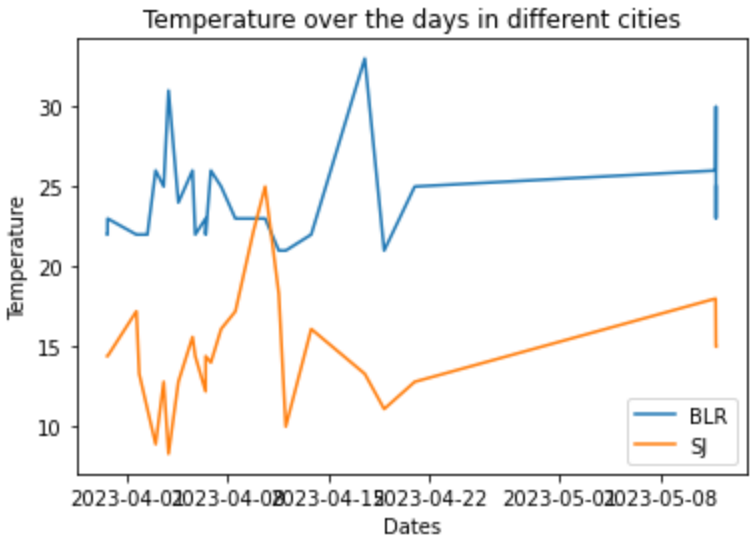
Weather data - I have used the weatherapi website to fetch the weather details on a more frequent basis for which the most recent report available is the current weather. Useful information in these reports includes temperature, pressure, wind speed, weather condition (e.g. rain, snow, sunny, clear), time, and precipitation. Additional metrics such as humidity and solar radiation can be derived from this report. The climate is a composite of weather conditions taken over a period of time and is connected to the tweets based on the time. So, each tweet has corresponding weather features attached to it.

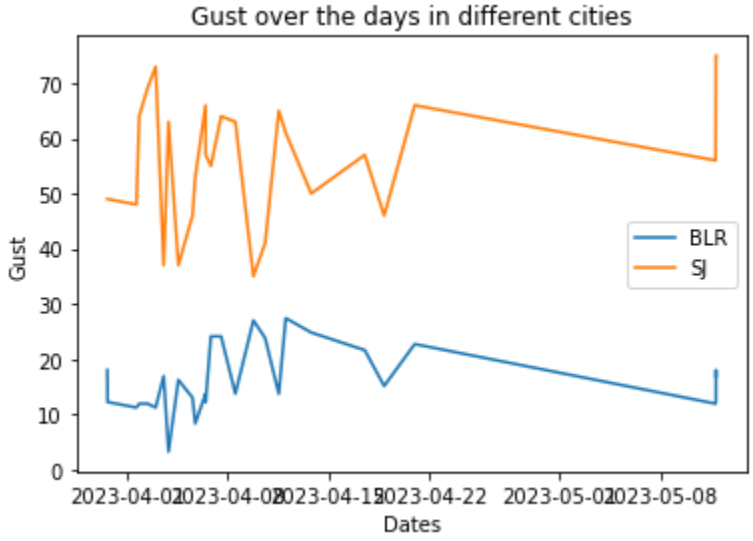
4. Analysis

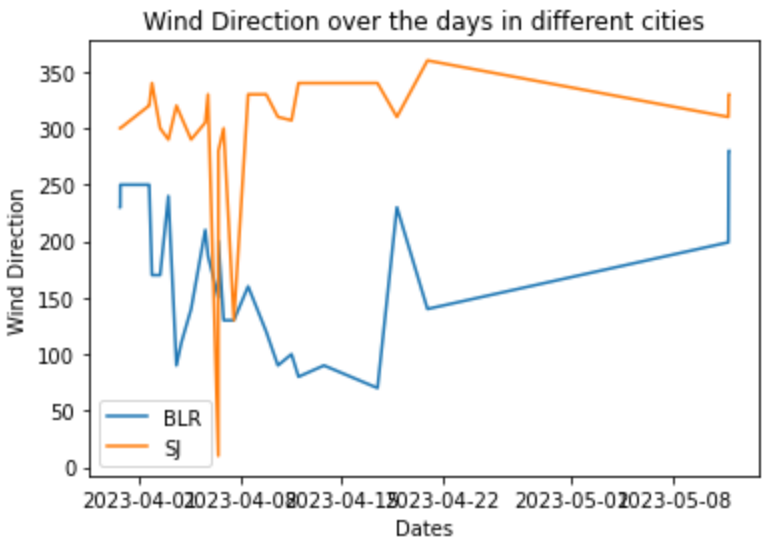
4.1. Exploratory Data Analysis



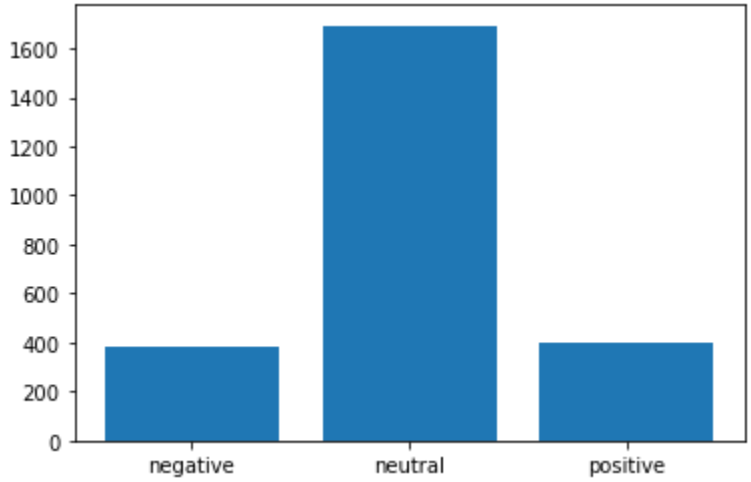
- We can see that the favourite count and retweet count are very less.
- The latitude and longitude are from the same range which is correct.
- Positive and negative probabilities have a majority of their values less than 0.4 whereas neutral has values more than 0.4
- Coming to the weather features they are evenly spread across the range with the exception of precipitation, vis and uv.

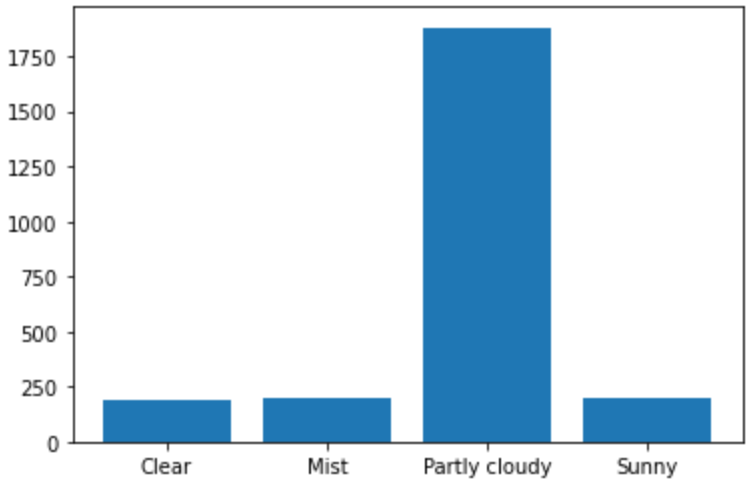

Temperature over the days in different cities
Temperature
Dates
BLR
SJ
30
25
20
15
10
2023-04-01
2023-04-08
2023-04-15
2023-04-22
2023-05-01
2023-05-08


Gust over the days in different cities
Gust
Dates
BLR
SJ
70
60
50
40
30
20
10
0
2023-04-01
2023-04-08
2023-04-15
2023-04-22
2023-05-01
2023-05-08

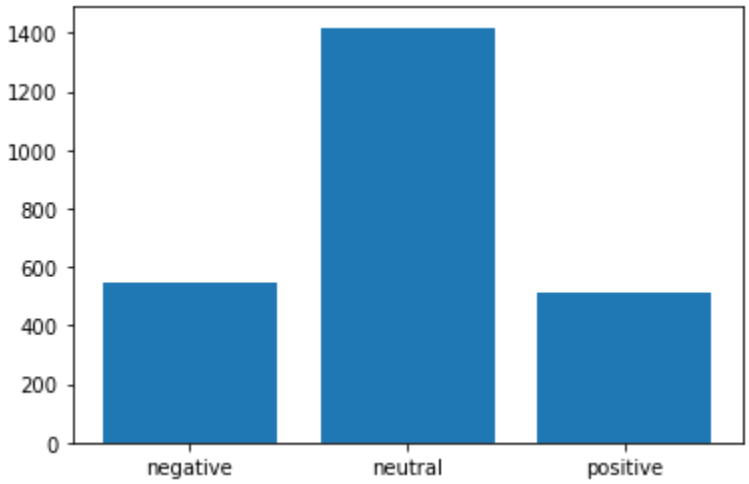

Wind Direction over the days in different cities
Wind Direction
Dates
BLR
SJ
0
50
100
150
200
250
300
350
2023-04-01
2023-04-08
2023-04-15
2023-04-22
2023-05-01
2023-05-08

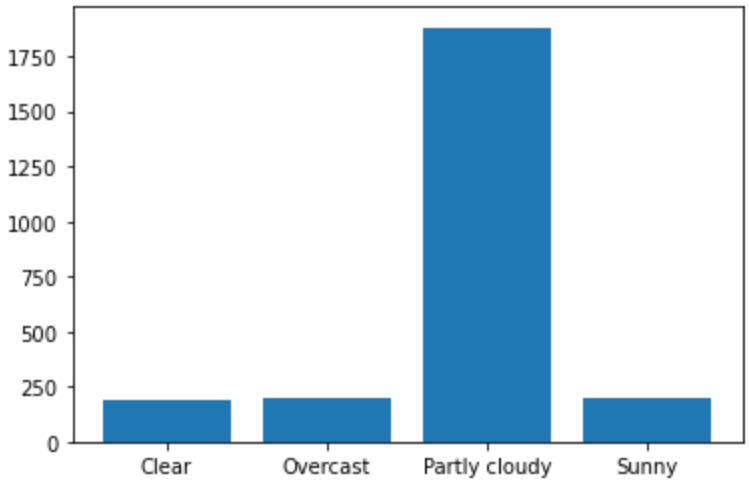
BLR

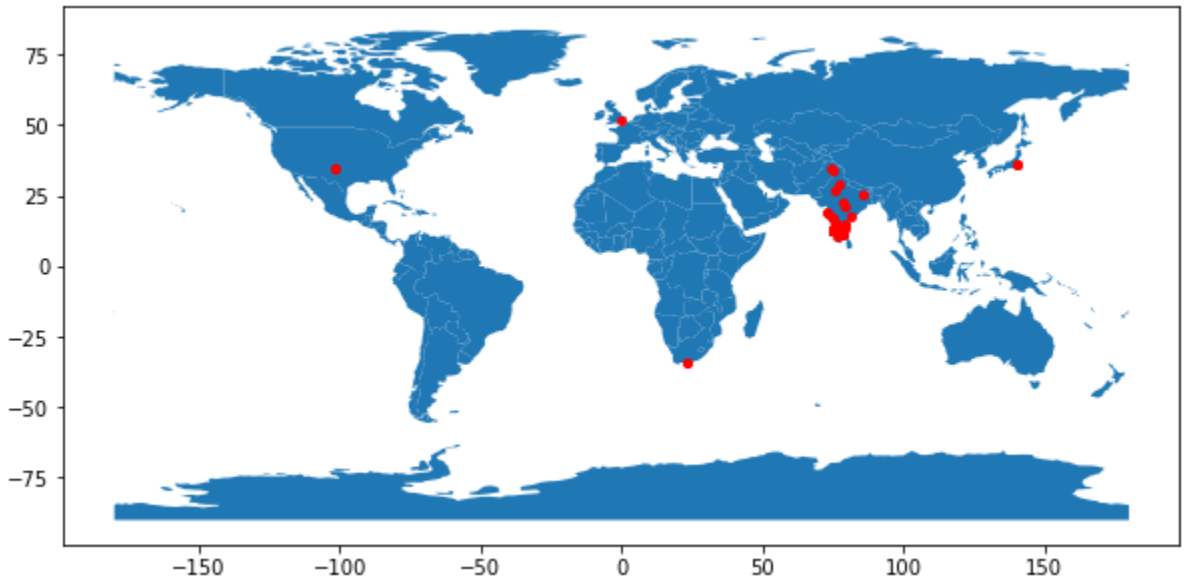

0
200
400
600
800
1000
1200
1400
1600
negative
neutral
positive

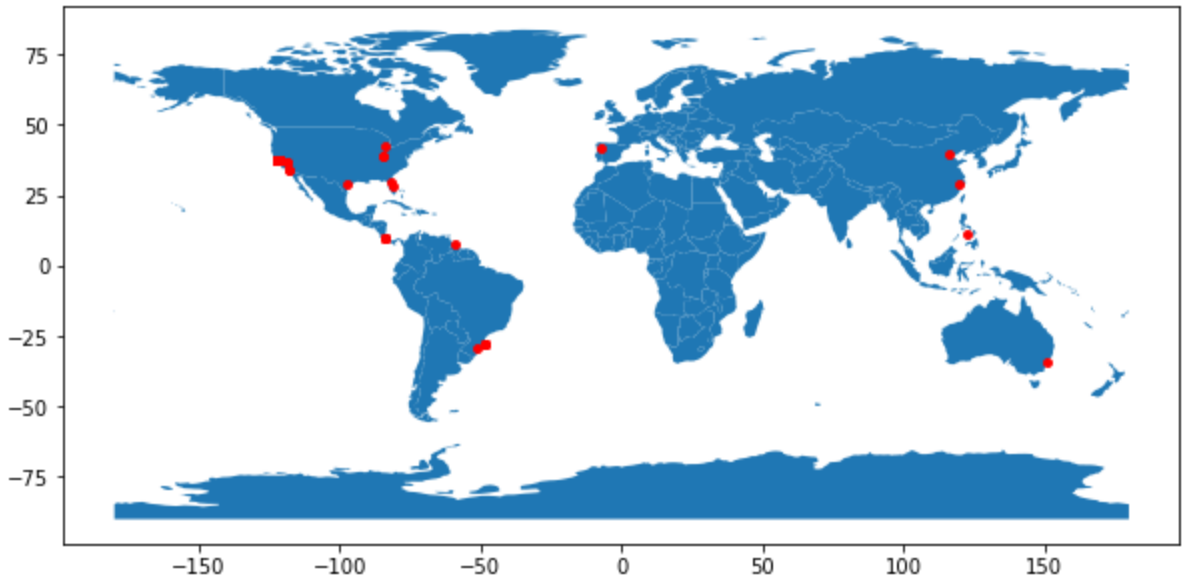

1750
1500
1250
1000
750
500
250
0
Clear
Mist
Partly cloudy
Sunny

SJ


1400
1200
1000
800
600
400
200
0
negative
neutral
positive


1750
1500
1250
1000
750
500
250
0
Clear
Overcast
Partly cloudy
Sunny


75
50
25
0
−25
−50
−75
−150
−100
−50
0
50
100
150



4.2. Proposed Methodology

The implementation plan involves the following steps:

1. Collecting social media data and weather data.
2. Preprocessing the data to remove noise and irrelevant information.
3. Performing sentiment analysis using the cardiffnlp/twitter-roberta-base-sentiment model.
4. Analyzing the results to determine the correlation between weather conditions and sentiment expressed in social media data using machine learning models.

4.2.1. Pulling tweets and weather data

We first take tweets from Twitter using Twitter API and extract only the necessary features like the tweet, language, location and many more and then perform some pre-processing like removing stopwords, removing punctuation, and many more. Then apply the Roberta model to find the sentiment of the tweets. On the other hand, we take in the weather using the weather API

and take in features like time, temperature, humidity, precipitation and many more. In the end link both the tweets and the weather using the time. This combination is given to the next step.

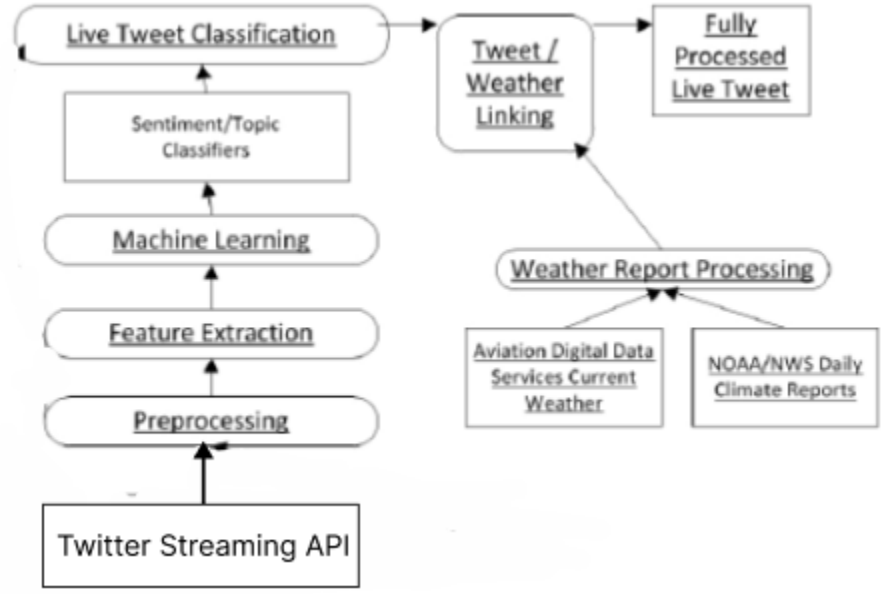


Fig. 1: Pulling and weather data

4.2.2. Pre-processing

There are a lot of pre-processing steps involved when working with text data. These include sentence segmentation, word tokenization, lowercasing, stemming or lemmatization, stop word removal, and spelling correction but are based on what the requirements are for the application.

These are a few steps we will be using:

a. Feature Extraction: We first do feature extraction of only the important features of the tweets like the tweet, location, language, time, and many more. Then we filter only the English tweets from them and only tweets limited to the United States. We could limit to only the United States but take in all languages and use a google translate API to convert the other

languages to English as we get to know the language of the tweets from the API. Then we perform the text cleaning steps which are described in detail next.

b. Change Case: Changing the case involves converting all text to lowercase or uppercase so that all word strings follow a consistent format. Lowercasing is the more frequent choice in NLP software.
c. Stop-Words Removal: Stop words are frequently occurring words used to construct sentences. In the English language, stop words include is, the, are, of, in, and. For some NLP applications, such as document categorization, sentiment analysis, and spam filtering, these words are redundant, and so are removed at the preprocessing stage.

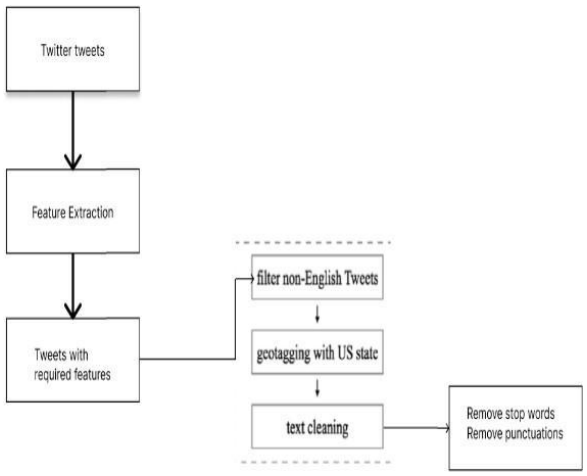


Fig. 2: Pre-processing tweets

4.2.3. Sentimental Analysis

For finding the sentimental analysis we would be using the twitter-roberta-base-sentiment model.

Architecture/Design:

The proposed methodology involves collecting weather-related and social media data and performing sentiment analysis with the twitter-roberta-base-sentiment model. Twitter-roBERTa-base is based on the Transformer architecture, which is a type of neural network architecture commonly used for natural language processing tasks. It consists of multiple layers of self-attention and feed-forward neural networks. The collected data is preprocessed and fed into the model, which determines whether the sentiment expressed in the text is positive, negative, or neutral. The model returns the probability of each label using a softmax function and then selects and returns the label with the highest probability.

Possible Solutions:

The proposed methodology can help in understanding the impact of weather on human emotions and behaviour by analyzing the sentiments expressed in social media data. It can provide insights into how people feel about different weather conditions and how it affects their moods and behaviour. This information can be used to develop strategies to improve well-being and productivity.

Advantages:

The proposed methodology is a low-cost method of analyzing the relationship between weather and sentiments expressed in social media data. It requires less computation and fewer resources to find results. It also allows for the quick and efficient collection and analysis of data from a large population. It can help in the identification of patterns and trends in human behaviour in relation to weather conditions.

Disadvantages:

The proposed methodology relies on the quality and quantity of data collected from social media platforms. The sentiment analysis model used may not be accurate in certain situations. There also might be ethical concerns related to the collection and use of social media data.

Feasibility:

The proposed methodology is feasible as it involves using readily available tools and techniques for sentiment analysis. The cardiffnlp/twitter-roberta-base-sentiment model is a

pre-trained model and can be easily accessed and used. The data collection process can also be automated using web scraping tools.

4.2.4. Finding correlation

We get the combined tweets from the above part and then perform some pre-processing like encoding etc. and do some feature extraction. In the end, find the correlation between the sentiment and weather features. We have to do this for each feature as each weather feature will have a different effect.

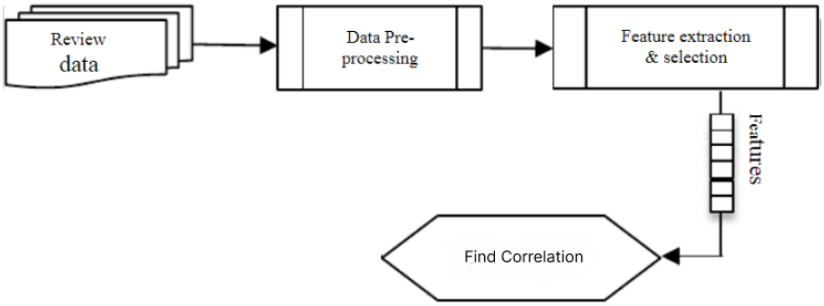


Fig. 3: Finding a correlation between weather and sentiment

4.3. Models

To find the correlation between weather and sentiment I would be using multiple machine models like decision tree, XGBOOST, Random Forest, Neural Networks, and Ensemble model. I will be applying all these techniques on two datasets i.e. tweets based on Bangalore and San Jose. In the end, I will be evaluating and comparing all the models using metrics like accuracy, precision, recall, and F1-Score.

The models I have used are mainly used for classification as the main objective of this project is to classify the sentiments based on the weather patterns.

4.3.1. Decision Tree

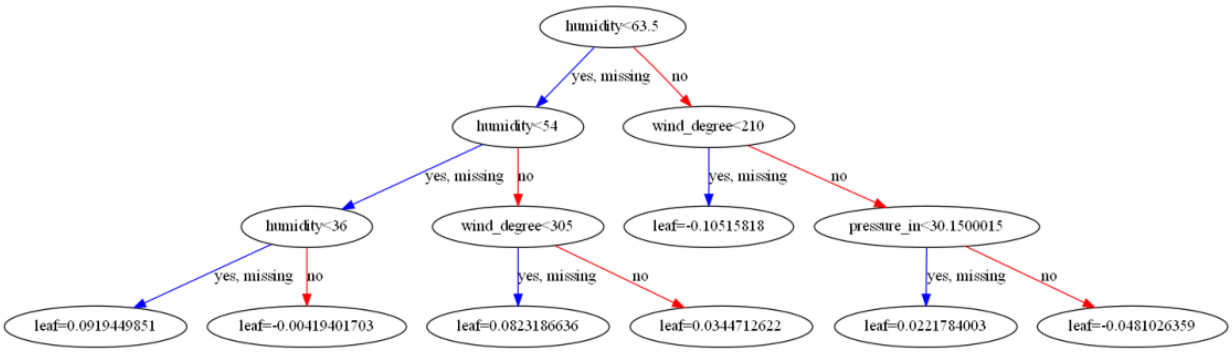
The decision tree is used for classifying the data based on the criteria it finds the most important. It can also handle categorical data which is very helpful.
In this project, I am using a decision tree to find the sentiment based on weather details and find the weather condition based on the previous sentiment probability for both the datasets.

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.
This is one of the decision tree models which was used in predicting the weather conditions.
Then I also used grid search for hyperparameters tuning which gives a slightly better result when compared to the base model.

4.3.2. XGBoost

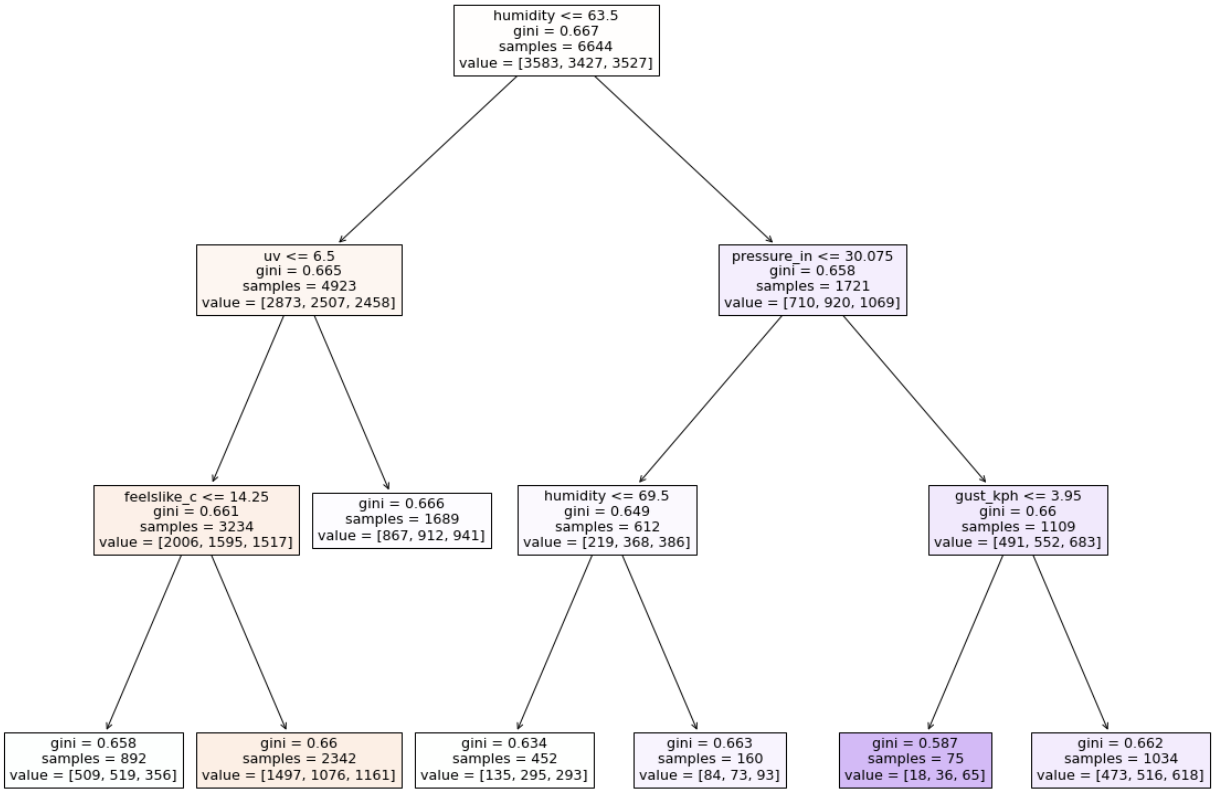
XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

For using XGBOOST I had to convert the categorical data to numeric as they don't work with categorical data. They are slower compared to decision trees as they have a much more complex architecture.



4.3.3. Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.

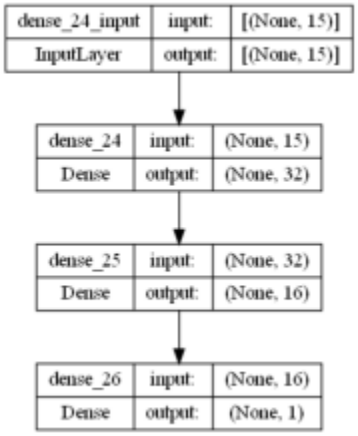


4.3.4. Neural Networks

Neural networks, also known as artificial neural networks (ANNs) or simulated neural networks (SNNs), are a subset of machine learning and are at the heart of deep learning algorithms.

Artificial neural networks (ANNs) are comprised of a node layer, containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network.

I am using neural networks for this because it can find some unseen correlations in the data and can be easily trained and hyperparameter tuning could be easily done. I am using 4 layers, first a sequential layer and then followed by 3 dense layers. In the 2nd and 3rd layer I have used relu as the activation function and a sigmoid function for the last layer. I am also using binary cross entropy as the loss function and adam as the optimizer.



4.3.5. Ensemble Model

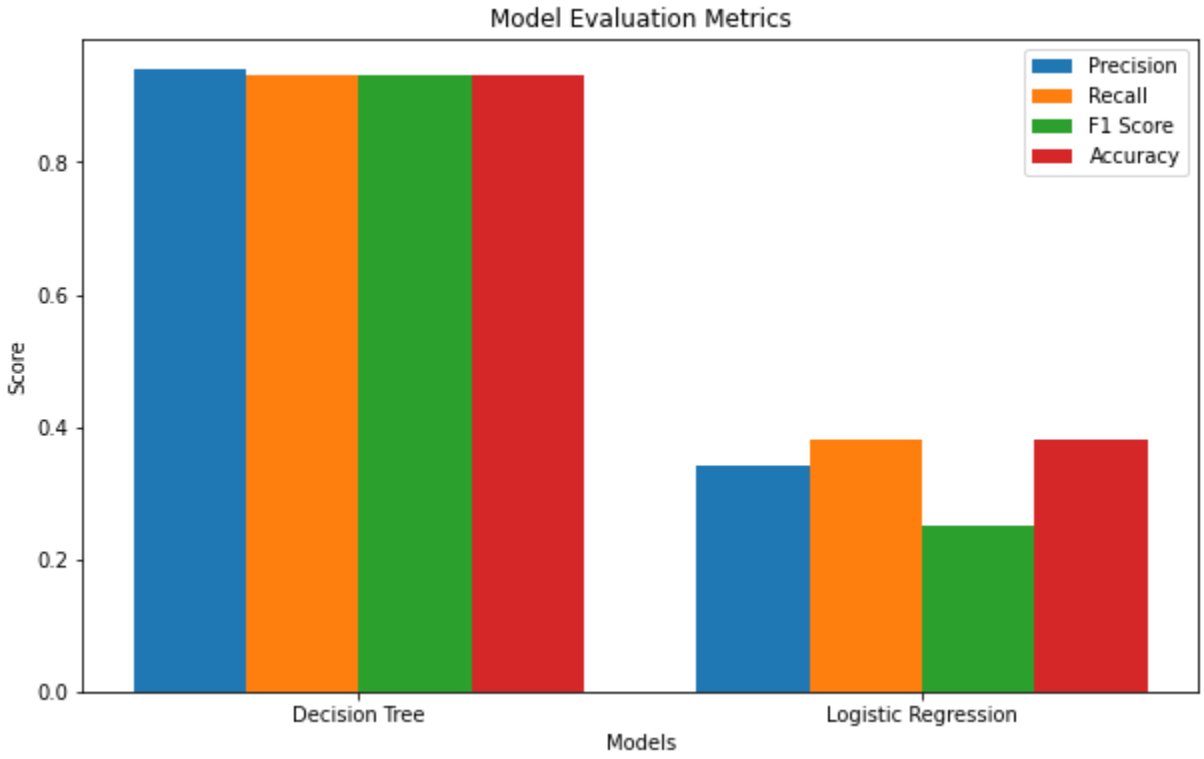
Ensembling involves combining all the models and getting a better result from the output of all the models. Here I am using all the above models discussed above and combining them and using the output from those models and sending the ensemble model to get a better understanding of the data.

5. RESULTS

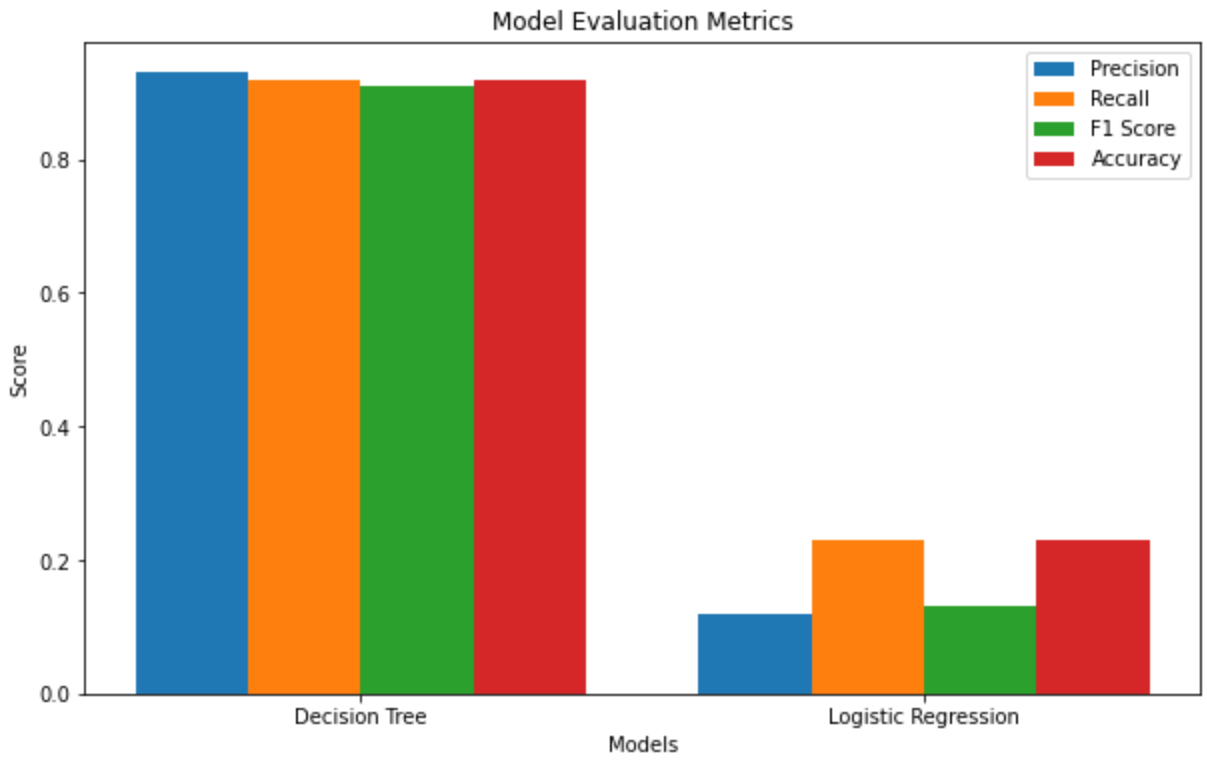
Predicting weather conditions based on the sentiment

When looking at the results we can see that both the datasets have very similar scores. But the Bangalore dataset performs better compared to the the San Jose dataset in both models.

Bangalore



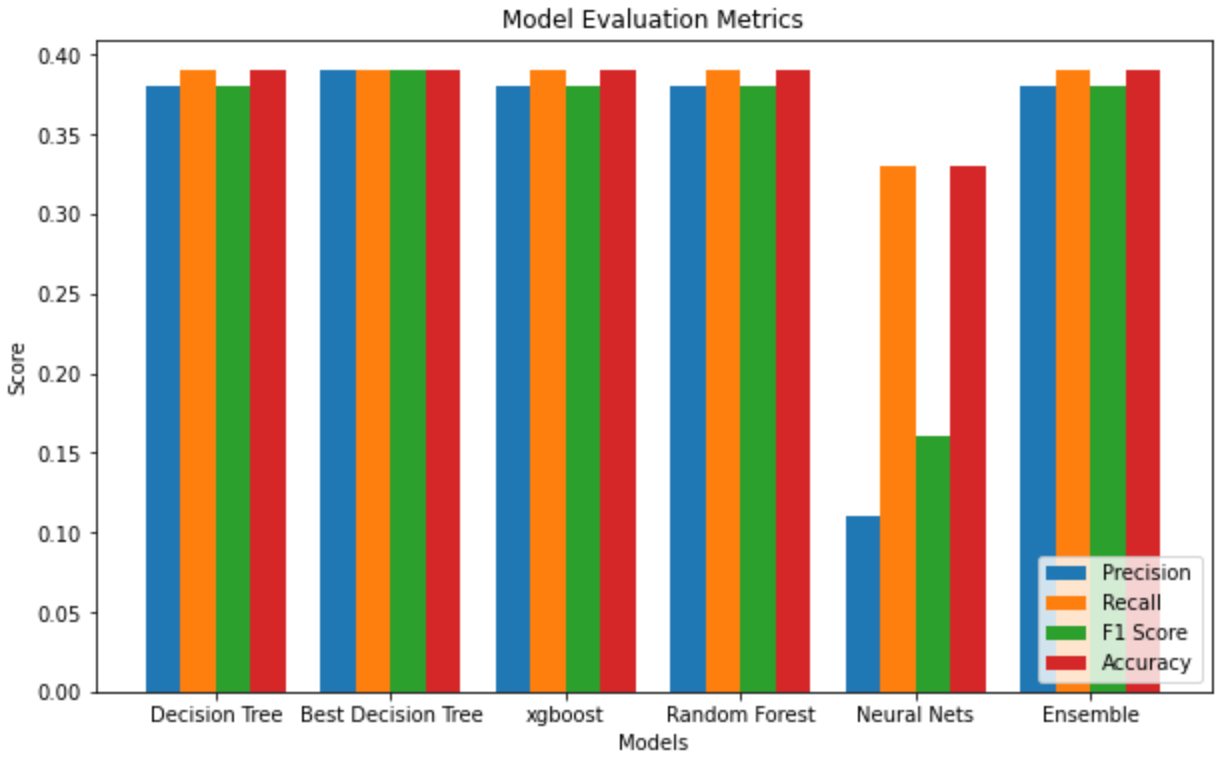
San Jose



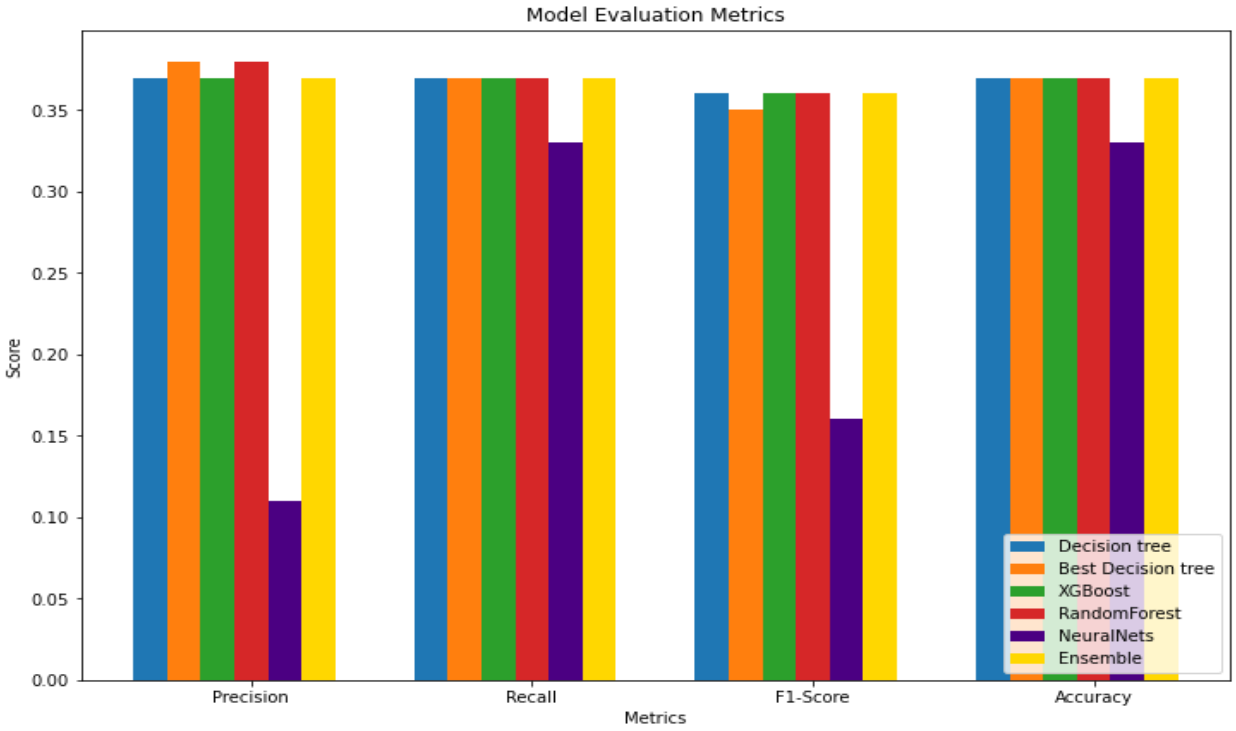
Finding the sentiment based on the weather patterns

When we look into the results we can see that the score of all models does not cross 40 for any metric. We can also observe that all the tree models perform very similarly and the neural networks perform the worst out of the lot.

Bangalore


Model Evaluation Metrics
Score
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Decision Tree
Best Decision Tree
xgboost
Random Forest
Neural Nets
Ensemble
Models
Precision
Recall
F1 Score
Accuracy

San Jose


Model Evaluation Metrics
Score
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Precision
Recall
F1-Score
Accuracy
Metrics
Decision tree
Best Decision tree
XGBoost
RandomForest
NeuralNets
Ensemble

6. Discussion

The proposed project aims to investigate the correlation between weather conditions and sentiment expressed in social media data. By analyzing tweets and their associated weather data, the project explores how different weather conditions affect people's emotions and behaviors. The use of social media data for sentiment analysis offers valuable insights into individuals' perceptions and experiences related to weather. The project leverages machine learning models such as decision trees, XGBOOST, random forests, neural networks, and ensemble methods to uncover the relationship between weather and sentiment.

One of the challenges in sentiment analysis across multiple languages is the variation in grammar, syntax, and cultural differences. By filtering tweets in English and using translation APIs for other languages, the project addresses the issue of multiple languages and aims to capture sentiments expressed in different languages. This allows for a more comprehensive analysis of weather-related sentiments across diverse linguistic communities.

Irony and sarcasm pose significant challenges in sentiment analysis, as they often involve the opposite meaning of what is

being said. Detecting irony and sarcasm accurately is crucial for understanding sentiment correctly. While the specific techniques and approaches for handling irony and sarcasm are not explicitly mentioned in the project description, it is an important aspect to consider during the sentiment analysis process.

Ambiguity and subjectivity in language present additional complexities in sentiment analysis. Words can have multiple meanings and can be interpreted differently based on context and individual perspectives. To mitigate these challenges, the project utilizes the roBERTa-base model, a transformer-based model trained on a large dataset of tweets, to capture the contextual nuances and improve sentiment analysis accuracy. However, the project does not provide detailed information on how the roBERTa-base model is specifically used to address ambiguity and subjectivity.

The project's objective is to determine the correlation between weather conditions and sentiments expressed in social media data. By applying various machine learning models, such as decision trees, XGBOOST, random forests, neural networks, and ensemble methods, the project aims to uncover patterns and relationships between weather features and sentiment labels. This analysis can

provide valuable insights into how weather impacts human emotions and behaviors, enabling the development of strategies to enhance well-being and productivity.

7. Evaluation and Reflection

The project's evaluation involves the performance assessment of the different machine learning models used for sentiment analysis and correlation analysis. Metrics such as accuracy, precision, recall, and F1-score are mentioned as evaluation criteria for comparing the models. By evaluating the models on two datasets based on different locations (Bangalore and San Jose), the project assesses the models' ability to generalize across different contexts.

The choice of evaluation metrics provides a comprehensive assessment of the models' performance in sentiment classification and correlation analysis. Accuracy is a common metric that measures the overall correctness of the model's predictions. Precision and recall provide insights into the model's ability to correctly identify positive, negative, and neutral sentiments. The F1-score is a harmonic mean of precision and recall, capturing the balance between the two metrics.

The evaluation process should also consider the limitations and challenges associated with sentiment analysis, including the inherent subjectivity of sentiment labeling and potential biases in social media data. It is crucial to interpret the evaluation results in the context of these limitations and consider the potential impact on the project's findings and conclusions.

8. References

[1] S. Golder and J. Macy, “Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures,” *Science*, vol. 333, no. 6051, pp. 1878-1881, 2011.

[2] S. Zimmerman and U. Kruschwitz, “Speaking of the weather: Detection of meteorological influences on sentiment within social media,” *9th Computer Science and Electronic Engineering (CEEC)*, Colchester, UK, 2017, pp. 1-6, doi: 10.1109/CEEC.2017.8101590.

[3] H. Li, Z. Jadidi, J. Chen, and J. Jo, “The Use of Machine Learning for Correlation Analysis of Sentiment and Weather Data,” in *Robot Intelligence Technology and Applications, RiTA 2017, Advances in Intelligent Systems and Computing*, vol 751, pp. 291-298, Springer, Cham.

[4] R. Anisha, R. Archana, and C. Niranjan, “Sentiment Analysis of Real Time Twitter Data Using Big Data Approach,” *2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*, pp. 1-6, doi: 10.1109/CSITSS.2017.8447656.

[5] S. Dutta, S. Ray, and S. Roy, “Correlation between weather and weather-related tweets — A preliminary study,” *2016 IEEE International Conference on Big Data (Big Data)*, Washington, DC, USA, 2016, pp. 3971- 3973, doi: 10.1109/BigData.2016.7841079.

[6] D. Petrova and V. Bozhikova, “Random forest and recurrent neural network for sentiment analysis on texts in Bulgarian language,” *2021 International Conference on Biomedical Innovations and Applications (BIA)*, Varna, Bulgaria, 2022, pp. 66-69, doi: 10.1109/BIA52594.2022.9831326.

[7] M. D. Albayrak and W. Gray-Roncal, “Data Mining and Sentiment Analysis of Real-Time Twitter Messages for Monitoring and Predicting Events,” *2019 IEEE Integrated STEM Education Conference (ISEC)*, Princeton, NJ, USA, 2019, pp. 42-43, doi: 10.1109/ISECon.2019.8881956.

[8] K. S. Vrunda and B. Jayasri, “Sentimental analysis of Twitter data and Comparison of covid 19 Cases trend Using Machine learn- ing algorithms,” *2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT)*, Bangalore, India, 2022, pp. 1-7, doi: 10.1109/GCAT55367.2022.9971980.

[9] K. S. Madhu, B. C. Reddy, C. Damarukanadhan, M. Polireddy, and N. Ravinder, “Real Time Sentimental Analysis on Twitter,” *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India, 2021, pp. 1030-1034, doi: 10.1109/ICICT50816.2021.9358772.

[10] S. C. Agrawal, S. Singh, and S. Gupta, “Evaluation of Machine Learning Techniques in Sentimental Analysis,” *2021 5th International Conference on Information Systems and Computer Networks (ISCON)*, Mathura, India, 2021, pp. 1-5, doi: 10.1109/ISCON52037.2021.9702430.

[11] D. Sharma, M. Sabharwal, V. Goyal, and M. Vij, “Sentiment Analysis Techniques for Social Media Data: A Review,” *First International Conference on Sustainable Technologies for Computational Intelligence*, pp.75-90, doi: 10.1007/978-981-15-0029-9 7.

-

[12] A. Goel, J. Gautam, and S. Kumar, “Real time sentiment analysis of tweets using Naive Bayes,” *2nd International Conference on Next Generation Computing Technologies (NGCT)*, pp. 257–216.

[13] Wikipedia contributors. Wikipedia.org. https://en.wikipedia.org/wiki/Main Page. (accessed May 07, 2023).

[14] B. O’Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, “From tweets to polls: Linking text sentiment to public opinion time- series,” *Proceedings of the 4th*

International AAAI Conference on weblogs and Social Media (ICWSM), 2010, pp. 122–129.

[15] J. Li, X. Wang, and E. Hovy, "What a nasty day: Exploring mood- weather relationship from twitter," in *CIKM*, 2014, pp. 1309–1318.

[16] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135,2008.

[17] ExchangeScale.com. https://exchange.scale.com/public/blogs/preprocessing-techniques-in-nlp-a-guide. (accessed May 07, 2023).

[18] F. Barbieri, J. Camacho-Collados, L. E. Anke, and L. Neves, "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 16-20 November 2020, vol. EMNLP 2020, pp. 1644–1650.