Final Project Report

Abinaya Janakan | A20376287

Prathyusha Melapindi | A20358277

Shringa Bais | A20382937

Tejas Sreenivasan | A20377089

Contents

INTRODUCTION

Psychological studies have shown that weather affects human lives, by having an impact on mood. Their study indicated that the seasonal weather changes make people experience aggressive, sad, lethargic, energetic mood changes throughout the seasons.

How would humans react if there was a change in the seasonal weather due to climate change? Observations held in the U.S have also shown that Climate change is evident. Scientific studies have shown that the average temperature of Earth's global surface has risen 1.5°F over the past few 100 years, and is expected to grow even further in the next few centuries. In this project, we aim to find analytical evidence on how weather conditions modulate human behavior and see how the impact of climate change affects their sentiment towards it.

HYPOTHESIS

We have considered that both the hypothesis is true and have proceeded with the further analysis which is described below and the final conclusion has mentioned what has happened when the hypothesis considered is true.

Hypothesis-01:

There is relationship between human behavior and climate change

Hypothesis-02:

There is relationship between climate change and happiness.

To Prove Hypothesis-01:

The main goal to prove this hypothesis is that when there is increase in population, we believe that there is increase in the amount of carbon dioxide in the environment which makes the temperature warmer over the years. So, we can prove the hypothesis by saying the increase in population changes the climate which is proved that human behavior is related to change in climate.

To prove Hypothesis-01, we have taken the data for the country India from the given url https://data.gov.in/catalog/global-average-temperature-and-atmosphere-concentration-carbon-dioxide. The data has the carbon-dioxide content over the years 2003 to 2013 in India.

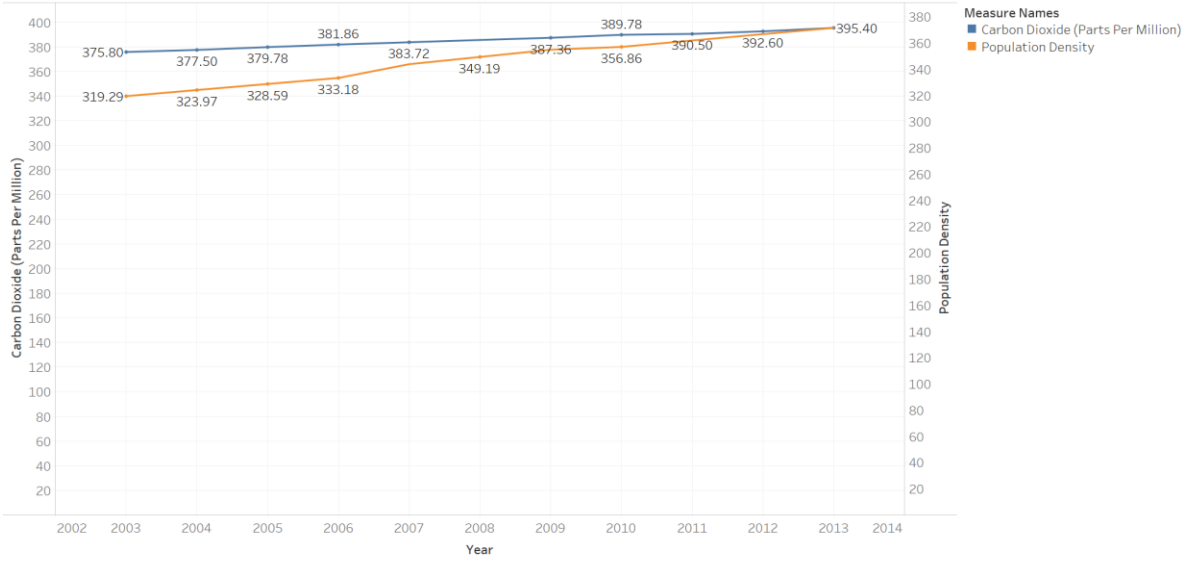
To get the population over the years 2003 to 2013, we have collected data from http://www.indexmundi.com/g/g.aspx?v=21000&c=in&l=en . This data has the population

over the years from 2003 to 2013 in the country India.

Below is the data for pollution and carbondioxide over 2003 to 2013

PollutionData.csv

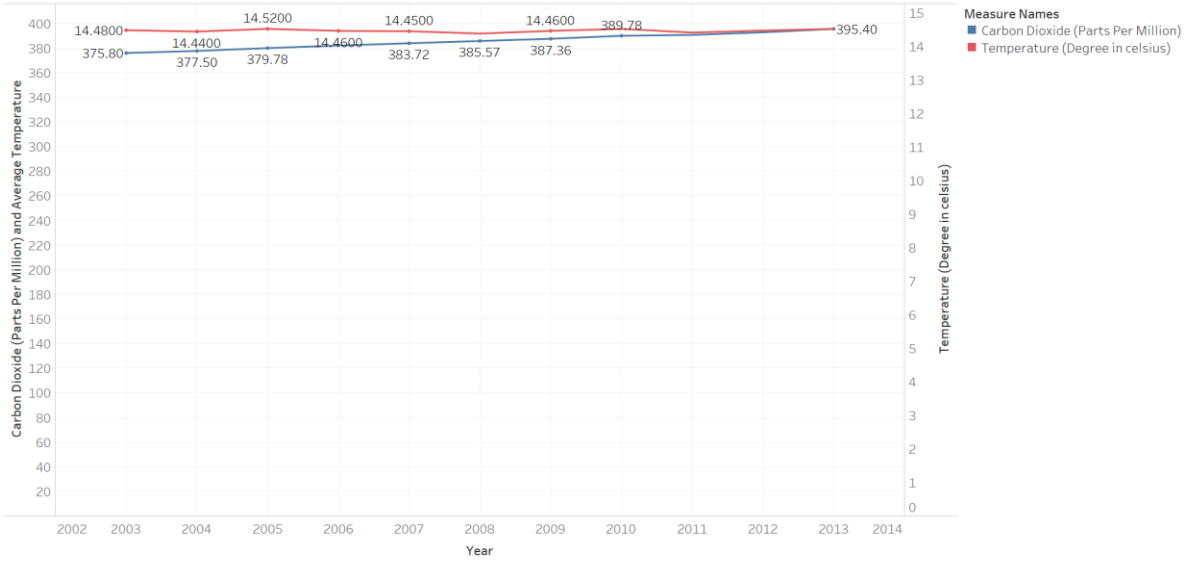
Sheet 1



The trends of Carbon Dioxide (Parts Per Million) and Population Density for Year. Color shows details about Carbon Dioxide (Parts Per Million) and Population Density.

The above graph clearly shows that if there is increase in the population then there is increase in the carbon dioxide content in the atmosphere. It clearly shows there is relationship between human behavior and climate change proving Hypothesis-01

Sheet 2



The trends of Carbon Dioxide (Parts Per Million) and Temperature (Degree in celsius) for Year. Color shows details about Carbon Dioxide (Parts Per Million) and Temperature (Degree in celsius).

To make the hypothesis-01 stronger we have also considered the average temperatures over the years 2003-2013 in India with in carbon dioxide in atmosphere which in the above graph shows that when there is increase in the carbon dioxide then there increase in the average temperature too.

To Prove Hypothesis-02:

To prove hypothesis-02 we have to consider the sentiments of the people because to get the happiness, we have to consider the social media where people express their feelings or concerns to get the happiness scores. This is done using Twitter analysis which is explained below.

DATA COLLECTION

Data Collection is the main phase of the analysis. The source of our data should be areas where humans can express their emotions against climate change. Many places where we can find human interactions with environmental activities are blogging sites, social networking sites, news articles, and social conferences. There are many other indirect ways of extracting human behavior such as public transport details, taxi usage, historic weather information etc. To maintain the integrity and quality of data, **Twitter** is used for data collection.

Below shown is the example tweets related to the keywords Global Warming and climate change.

Below is the sample tweets collected on the keywords for natural disasters

WHY TWITTER?

The source we refer to for this project is Twitter, a very popular social networking micro blogging site. A unique engagement is the messages (known as tweets) posted have a size restriction of 140 characters, which keeps things scan-friendly, simple and makes focused and clever use of language in every message. Twitter also provides REST API's for programmatically data extraction through their web services.

We use the twitter Streaming API to capture tweets posted at real time. This service was

run intermittently over the week and over 20,000 tweets were collected. We utilized the glossary of climate change terms to target tweets about environmental science. We have also analyzed the message contents to assess user attitudes towards climate change and sentiment in user interactions.

Below is the csv file of the tweets after data preprocessing of twenty thousand tweets which resulted to 3500 tweets.

Tweets.csv

KEYWORDS

The diction used to describe the climate change on Twitter varies among users. Example, some users may use the keyword ‘global warming’ and others may use the keyword ‘Climate change’.

There are also people who have used different hash tags to express their feelings towards climate change. Some of the keywords used to collect the tweets are ‘Abrupt ClimateChange’,’Anthropogenic’,’AtmosphericLifetime’,’BlackCarbonAerosol’,’CarbonCycle’, ’Ocean Acidity’, ’Climate Change’, ’Climate System (or Earth System)’,’Fluorinated Gases’, ’Global Average Temperature’, ’Global Warming’.

By considering the sample tweets, it is found that the tweets having the keyword ‘Global Warming’ has more negatively rated words than the keyword ‘Climate’ such as ‘die’, ’threat’, ’fraud’ etc. The words “disaster” and “hurricane” are used more frequently in climate tweets, suggesting that the subject of climate change co-occurs with mention of natural disasters, and strong evidence exists proving Twitter is a valid indicator of real time attention to natural disasters.

To extend the key words we had gone through the disasters list in United States and have taken few keywords such as Tsunami, Earthquake, Riverine floods, Storm and ozone layer. We have found tweets that people show a lot of concern when there is a natural disaster.

DATA CLEANSING

Data cleansing is done by detecting and removing the inaccurate records such as the value [RT] is removed from the tweets as it specifies the meaning that the tweet is a retweet, URL’s, emoticons, hash tags, special characters, and punctuations.

After the first stage of removing characters, we start by removing duplicate tweets.

We observed that, many of the tweets where programs re-tweet of popular awareness. By removing the repeating tweets, we could reduce the tweets from 20,000 to 3500.

This process is done repeatedly to remove the tweets which are about climate change but does not have much meaning in it by changing the glossary of keywords and running the analysis repeatedly.

SENTIMENT ANALYSIS

We train a Logistic classifier from a dataset of pre labeled tweets [(Sentiment140,n.d.)].Over 16 million tweets have been pre labeled with their sentiments, and this data is cleaned and loaded into our logistic classifier as training data. Using TF-IDF, we quantitatively construct a vector representation of the possible n-grams (tri grams in this case) from the collection of tweets to help better predict the sentiment of sentences.

“I like this product a lot, not as much as product X”. A trigram tokenization will help phrases like “not_as_much” qualify as negative word phrases. For this project, we will aim to use Uni-gram, Bi-gram and tri-gram tokenization for accurate sentiment prediction.

The Training data set was limited from 60 million to 190000 tweets for processing complexity reasons.

The training tweet's features were extracted using the TF-IDF Vectorizer utility from Python’s sklearn kit. After testing multiple configurations, we finalized on using the trigram tokenizer. Using the feature matrix containing TF_IDF features where then fit into a Machine Learning Model (Logistic regression here)

The tweets regarding climate change collected from the streaming API are featurized again with the tf_idf feature and then transformed for classification as part of the testing/classification phase

The training data found 175725 terms to featurize under a unigram tokenizer and has a training accuracy of 83.1%. [accuracy is expected to be high for a trigram tokenizer].The top weighted terms for the positive class as per the training data are: [('thanks', 7.28523), ('thank', 5.8425), ('welcome', 5.47061), ('great', 4.86411), ('followfriday',4.65275), ('awesome', 4.650827), ('worries', 4.3911631332831274), ('congratulations',4.33324), ('worry', 4.24242), ('congrats', 4.24109)]

The top weighted terms for negative class as per the training data are:

[('sad',-11.68110),('miss',-8.692),('sucks',-8.07944), ('sorry',7.7858151462349952), ('poor',-7.55605), ('wish',-6.88595), ('sadly', -6.46245), ('unfortunately', -6.313391), ('missed', -6.3042531),('bummer', -5.50835)]

An example of tweet classification is here below, with the probability values

estimated by the Logistic classifier. We are relying on a binary classification, being [Positive/Negative]

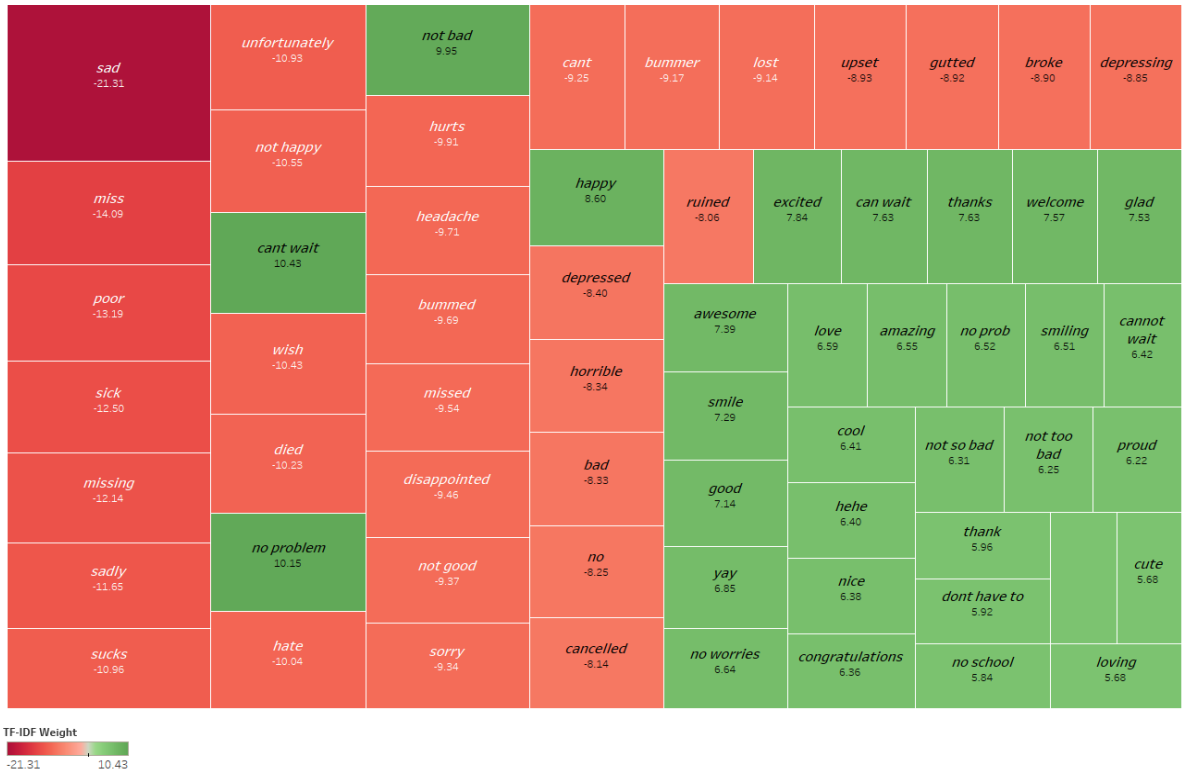
<u>One tweet from the Positive class is below:</u>

('hiene my of the future Thank you for denying climate change and helping to ruin the env ironment for the future generations', 'Positive', 0.88727462177988603, [('Negative',0.112 72537822011397), ('Positive',0.88727462177988603)])

<u>One tweet from the Negative class is below</u>:

('This is as ad news update but we have only ourselves to blame for all the climate change effects we see and fee', 'Negative', 0.93378491941835851, [('Negative', 0.9337849194 1835851), ('Positive',0.066215080581641481)])

TF-IDF



The above diagram shows the TF-IDF weight for the words in the tweets. As we can see the highest weight for the word ‘sad’ is -21.31 which is classified as a negative word. Similarly, the word ‘happy’ is given the highest positive score of 8.60.

INITIAL CONCLUSION

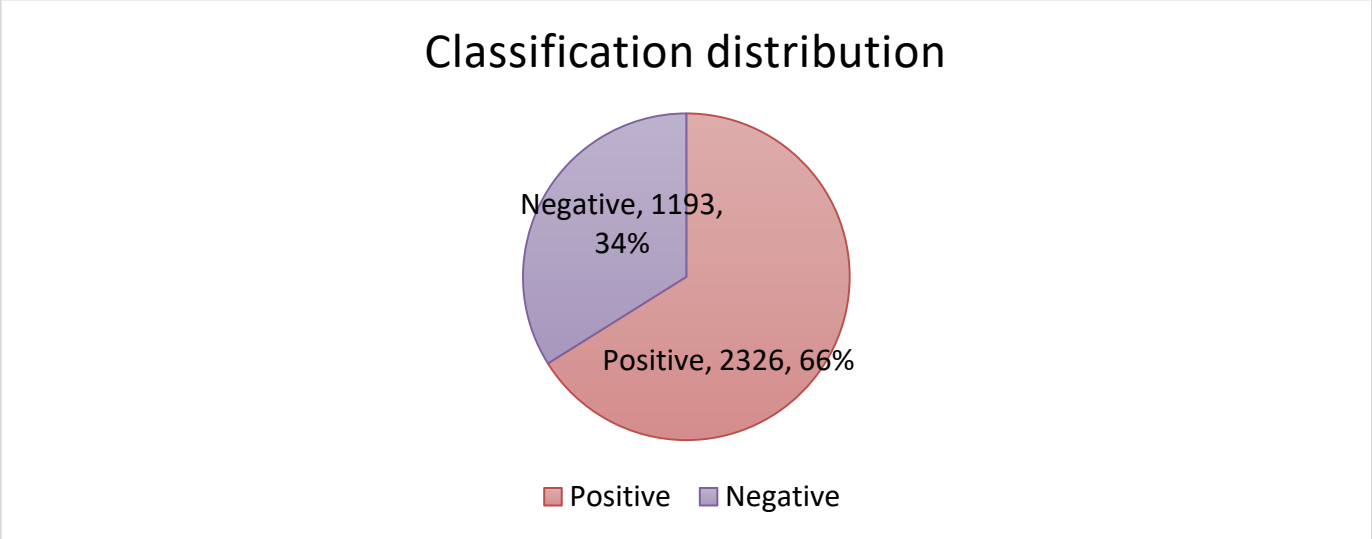
In the initial findings, we have done analysis for the tweets which have climate-related keywords and found three things

- Their concern with the change in climate.
- Their concern for the future generation.
- The sentiment of users across a geographic location where global warming awareness programs are held.

CHALLENGES:

1. We have found inconsistencies in our sentiment analysis with the use of just Uni-gram tokens. So we worked to include bi-gram and tri-gram tokens to get more accurate sentiment predictions from our classifier. Mean while we have also started collecting more tweets on a weekly basis through the twitter steaming service across various other climate change keywords.

2. The Challenges faced during the project are the tweets collected using the climate related keywords are mostly awareness programs which have classified into positive tweets. The challenge was to analyze that the tweets collected was related to the human behavior.



By seeing the above pie chart, the expectation was that there would be more number of tweets that are classified into negative rather than positive. But unfortunately the tweets majority of tweets were classified into positive rather than negative. That is when the projects scope had to be extended.

In addition to what has been done, we extended the analysis by collecting the natural disasters list in the past two years from http://emdat.be/disaster_list/index.html. These natural disasters list is used to collect the tweets from twitter. This is done because when natural disaster occurs then there would be people tweeting on twitter expressing

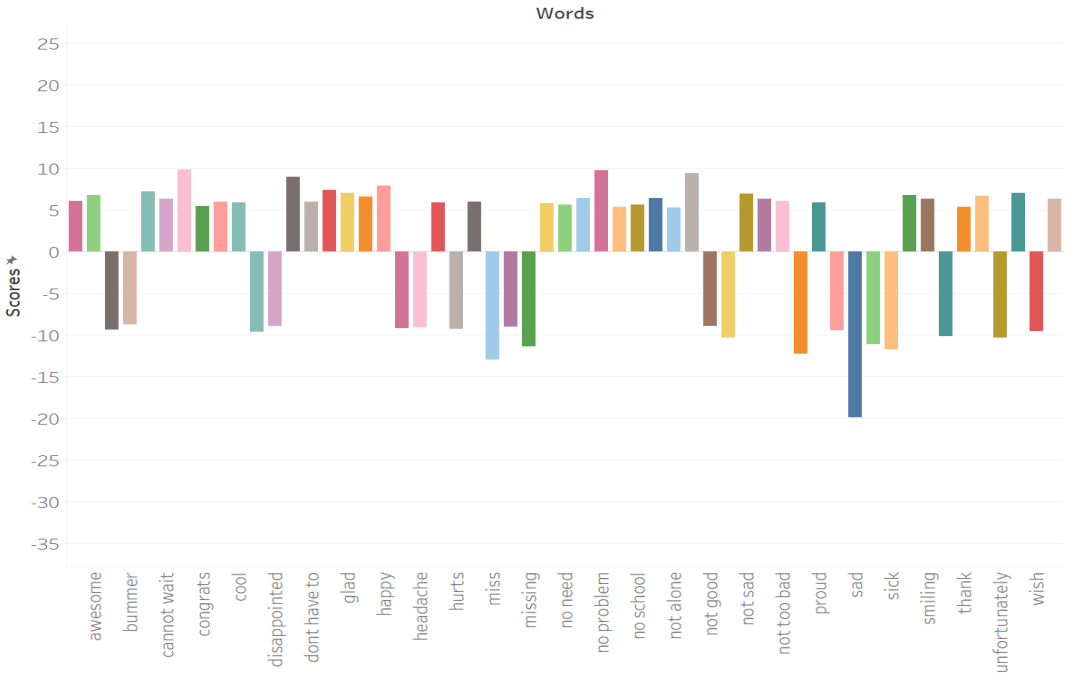
their concerns in the climate change. Although there may be awareness programs when the disaster occurs, people will show their concerns on the change. Sentimental Analysis is being performed on the tweets collected on natural disasters.

HAPPINESS INDEX:

Happiness index is measure by taking the cities data with considering so many factors such as GDP, Social support and healthy life etc. But considering happiness only on the way climate changes does not yield good results, because not all the people are happy due to change in climate change nor sad due to change in climate.

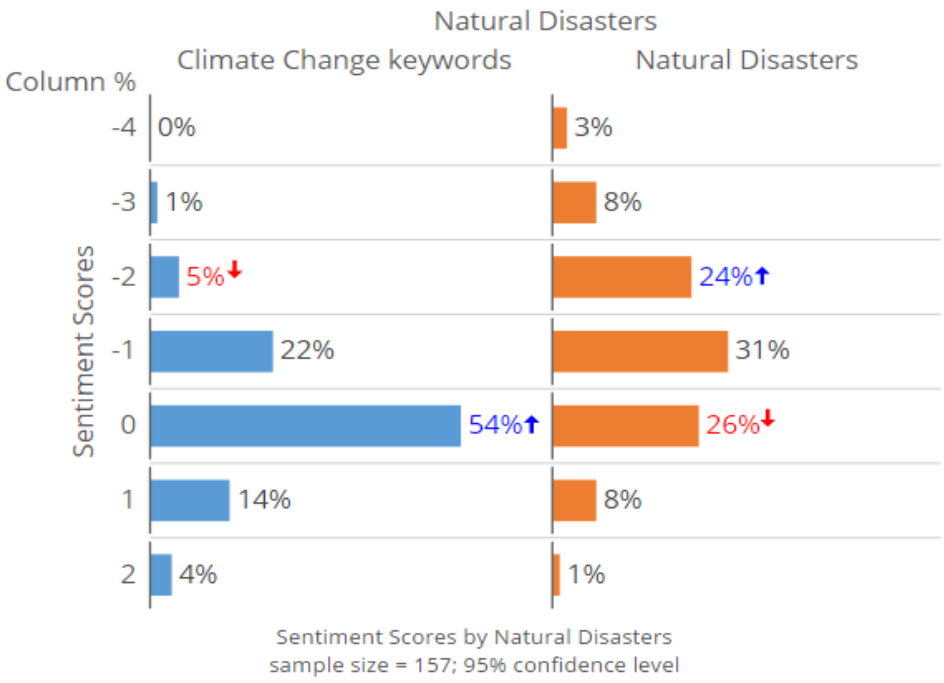
The **Hedonometer** is the tool that measures the happiness index on twitter every day. So, Hedonometer is implemented using python language which takes two inputs which are the list of tweets collected and the dictionary of words which has scores. The Hedonometer checks the words in the dictionary and checks through the tweets if there are matching words, based on the number of words matches it calculates the happiness index for that particular tweet i.e. Happiness score is calculated based on the average happiness score of the matched words contained within each tweet.

VISUALIZATIONS:



The above visualization shows the words with their scores with the score associated with it. The words taken are the bi-grams because we have found the uni-gram

words cannot yield good results to analyze the sentiment of the people.



In the above visualization is done by taking some collected tweets from both the climate related keywords and natural disasters. In the visualization the column % defines the sentiment scores calculated for all the tweets which is positive values, negative values and 0 is considered as neutral sentiment. For the Climate Change tweets the neutral percentage is high i.e. the value 0 has more than half which is 54%. As seen the tweets which are related to the climate change keyword has focused mostly on the awareness programs which just gives a neutral sentiment and the negative sentiments are all people concerns about climate change. The climate change tweets have more positive tweets too which can be said that the people are speaking well about the awareness programs and are giving positive comments.

On the other hand, the natural disaster tweets are having more negative sentiments and less positive sentiments. Which shows there is 26% decline in neutral sentiments too which says the people are sadder with the disasters.

This concludes the **hypothesis** which is the human's behavior change when the climates changes i.e. people are sadder when the climate is changing and the relationship between climate change and happiness is also shown in the above figure.

FINAL CONCLUSION

- Learned to apply Data Preparation and Analysis concepts on real world data and how it should be handled.
- When doing supervised learning makes sure you know the distribution of your classifiers. The distribution can create problems.
- Understanding the correct data for any analysis is very important.
- Human behaviors towards climate change increase in population and activities that lead to emission of carbon dioxide or methane gases are largely responsible for current trends in global climate change.
- Above analysis shows perfectly that any bad climate change affects human behavior badly and on the other hand making any change towards improving the environment condition also brings change in human behavior.
- Using happiness Index we have shown above in which areas people are more concerned towards climate change and taking various steps to improve it and make their surrounding better i.e. increasing happiness Index.

REFERENCES:

http://www.sciencedirect.com/science/article/pii/S0959378015000369

http://www.cse.scu.edu/~yfang/climate-fang.pdf

https://www3.epa.gov/climatechange/glossary.html

http://emdat.be/disaster_list/index.html

http://tweettracker.fulton.asu.edu/tda/TwitterDataAnalytics.pdf

http://hedonometer.org/index.html

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0026752

http://www.indexmundi.com/g/g.aspx?v=21000&c=in&l=en

https://data.gov.in/catalog/global-average-temperature-and-atmosphere-concentration-carbon-dioxide

Data Sources:

http://www.indexmundi.com/g/g.aspx?v=21000&c=in&l=en

https://data.gov.in/catalog/global-average-temperature-and-atmosphere-concentration-carbon-dioxide


tweets.xlsx


TweetScores.csv


Tweets.csv


tweets (2).csv


Sentimental_Analysis.xlsx


Disasters_tweets.xlsx


Pos_tweets.csv


Neg_Tweets.csv


Pos_tweets1.csv


Neg_Tweets1.csv


Pos_tweets2.csv


Neg_Tweets2.csv


Pos (1).csv


Neg (1).csv

Presentation slides:


Behavioral analysis.pdf