What are KOT saying about climate change

# **Business Overview**

## **Overview**

Climate is more than just one or two rainy days. Climate describes the weather conditions that are expected in a region at a particular time of year.

Climate change on the other hand describes a change in the average conditions — such as temperature and rainfall — in a region over a long period of time.

Global climate is projected to continue to change over this century and beyond. The magnitude of climate change beyond the next few decades depends primarily on the amount of heat-trapping gases emitted globally, and how sensitive the Earth’s climate is to those emissions.

Scientists say that even if we were to stop emitting greenhouse gasses completely, it is still expected that the average temperature will continue to rise.

The effects of climate change are More frequent and intense: droughts, storms, heat waves, rising sea levels, melting glaciers and warming oceans can directly harm animals, destroy the places they live, and wreak havoc on people's livelihoods and communities. As climate change worsens, dangerous weather events are becoming more frequent or severe.

The effects speak for themselves.

## **Objectives**

Our main objective for this study will be to investigate the trends and conversations that Kenyans on twitter are discussing about climate change.

Our main objective for this study will be to build a classification model that can classify text data on the impact of climate change as either famine, rain or air pollution related with an accuracy of at least 80%. **Change scoring method**

### Specific objectives

1. What are the main natural disasters that happen due to climate change?
2. To determine whether specific natural disasters happen during the same time period every year?
3. To determine whether conversations about climate change occur all through the year or do they peak on specific dates

To achieve the objective of this study, the analysis will seek to answer the following questions:

1. To classify tweets talking about natural disasters e.g Droughts, Floods, Forest fires.
2. Is there a trend on the number of tweets per day or month speaking about climate change?
3. Is there a trend on the number of tweets per day or month speaking about specific natural disasters?
4. To classify tweets talking about climate change precautions e.g Planting trees, alternative transportation, reduced carbon emissions, carbon offsetting.

## Impact

This research seeks to create a clear dashboard showing trends in conversations on twitter about climate change.

Our project intends to create a dashboard that will raise flags to the government whenever tweets about climate change and by extension specific natural disasters are under discussion by Kenyans on twitter.

E.g, perennial floods and droughts happen in Kenya around the same period every year - Kenyans on twitter speak about this and by extension climate change but the government doesn't seem to take notice.

**Our project intends to create awareness on the impact of climate change that is being experienced in the country.**

## **Data Mining Goals**

Our data mining goal for this project is to investigate the trends in conversations about climate change.

Potential questions for consideration include:

1. During what month do Kenyans speak most about climate change?
2. During what month do Kenyans speak most about Droughts?
3. During what month do Kenyans speak most about Famine?
4. During what month do Kenyans speak most about how to deal with climate change?
5. What method of climate change prevention is most talked about by Kenyans on twitter
6. What natural disaster is talked about most - possibly per year

## **Success Criteria**

A successful analysis will provide insights on the trends in climate change conversations in Kenya and a text classification accuracy of at least 80%

## Assessing the situation

**Assumptions**

There were no paid advertisements.

**Resource inventory**

1. Datasets :
2. Softwares: Github, Google collab, tableau **Plotly dash, streamlit, deployment**

**Implementation plan**

| Phase | Time-Frame |
| --- | --- |
| Formulation of Research Question | 30 minutes |
| Business Understanding | 1 hour |
| Data Understanding | 1 hour |
| Data scraping | 4 hours |
| Data Preparation and Cleaning | 5 hour |
| Data Analysis | 1 hours |
| Summary and Conclusion | 2 hours |

# **Data Description**

For this analysis, the data available has 14 columns and 4,895 rows collected from 2015 to 2020. The information available in the dataset is described in the table below:

| **COLUMN** | **DESCRIPTION** |
| --- | --- |
| 'user', | Unique identifier of the user who tweeted or retweeted |
| 'tweet', | The actual tweet |
| "location", | The location of the user tweeting |
| 'description' | Description of the tweet |
| ,'friends\_count', | People who the tweeter is connected with |
| 'followers\_count', | People who follow the tweeter |
| 'statuses\_count', |  |
| 'created\_at', | The date and time |
| 'retweet\_count', | Number of retweets |
| 'hashtags' | What hashtags does the tweet have |

# Data Preparation

* **Scraping and reading the datasets**

We first installed and imported the required libraries to help in the analysis. After, we scraped and loaded our dataset and created a dataframe. We then previewed both head( first 5 rows) and tail( last 5 rows) to get an understanding of the information we shall be analysing.

* **Exploring the data**

Afterwards, we checked the shape of our dataset. We checked the for the information of each column, to understand the datatypes of each, as well as a statistical summary of the data.

* **Data cleaning**

1. **Uniformity**: We checked for the uniformity in naming the columns names and removed any white spaces in the our columns as well as renaming our column names to lowercase.
2. **Missing Values**: checked for any null value in our data. Fortunately for us we found that our data had no missing values.
3. **Duplicates**: We checked if there were any duplicates in our dataset. Fortunate enough, our data contains no duplicates
4. **Irrelevant Columns**: Column names are irrelevant for our analysis, hence they were dropped.
5. **Outliers**: We decided to check for outliers in our dataset. Our dataset had a lot of outliers. We decided not to drop them as dropping them would affect the accuracy of the data analysis leading to results being inconclusive and incorrect.