NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR

Cachar, Assam

B.Tech. VIth Sem

Subject Code: CS-331

Subject Name: Social Network Analysis

Topic:

Geospatial Sentiment Analysis

Submitted By:

Students of Computer Science and Engineering:

1. Bhargab Nath - 1912019
2. Hrishikesh Dutta - 1912057
3. Pratik Gupta - 1912077
4. Rahul Gautam Singh - 1912082
5. Subhojit Ghimire - 1912160

Under the Guidance of:

Dr. Anupam Biswas

Assistant Professor

Department of Computer Science and Engineering

Contributions

1. **Bhargab Nath:**
    a. Word Cloud & Conclusion
    b. Data optimisation

2. **Hrishikesh Dutta:**
    a. Sentiment calculation

3. **Pratik Gupta**
    a. Map visualisation
    b. Dataset preparation

4. **Rahul Gautam Singh**
    a. Data refactoring
    b. Node clustering

5. **Subhojit Ghimire**
    a. Dataset Preprocessing
    b. Report Preparation

Geospatial Sentiment Analysis

ABSTRACT:

With the increase in global trend of social media usage, it is becoming immensely easy to keep oneself updated on the global happenings. From happy occasions to sad incidents, the information of it all is on social media platforms like Twitter, Facebook, Instagram, etc. If these available data are used correctly, we can train a model to give us a brief overview on happenings around the world and analyse whether these happenings are positive or negative to an individual or a community as a whole. Our aim in this study is to analyse the tweets geospatially and then perform geo-spatial sentiment analysis based on the location and learn if people in a certain location are experiencing happiness or sadness. We will analyse the entire sentences, pictogram, logogram, ideogram and hashtags to train our model into figuring out whether these data carried positive quality, negative quality or neutral sentiment.

INDEX TERMS: Geospatial, Sentiment Analysis, Word cloud, Twitter dataset

INTRODUCTION:

Sentiment analysis is a process of examining provided data, usually text data, to extract the hidden sentimental data in order to study the provided generic data in broader terms and understand the equation better. Wikipedia defines sentiment analysis as the use of natural language processing (NLP), text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information.[1] Geospatial refers to geography and mapping. MAPS defines geospatial information as 'place based' or 'locational' information. [2]

In whole, geospatial sentiment analysis is the practice of mapping sentimental data to model and represent how people, objects and phenomena interact within space, as well as to make predictions based on trends in the relationships between places.[3] Usually, the data in such analysis are collected through customers' textual feedback via surveys to analyse people's opinion. In other cases, the data are imported from the online public media to get insights from any written text.

This study involves analysing people's tweets as data from Twitter to understand people's mood and finding interesting location-based patterns.

Furthermore, the results obtained in this study were plotted in the world map and the effects of clustering and modularity of the so-obtained network were graphically studied. Interestingly, in some cases we can observe unusual patterns, while in some other areas, we can observe overlapping and high clustering with a lot of conflicting sentiments, which we will discuss later in the study.

METHODOLOGY:

For this study, we strictly adhered to python and its dependencies for analysing the data as well as for visually representing the so-obtained results.

We used and compared various different methods for sentiment analysis on tweets (a binary classification problem). The training dataset is expected to be a csv file of type Demojized, Captions, Latitude, Longitude, Created Time, Sentiment, Sentiment Score, Full text where the Demojized column consists of raw tweet text with the emojis translated, Captions consisting the captions of the instagram post mentioned in the tweet, Full Text consisting whole demojized and captions which are further translated to English, Sentiment score is between -1 (negative) to 1 (positive), and Sentiment is the remark given within a certain range of sentiment score.

There are some general library requirements for the project and some which are specific to individual methods. The general requirements are as follows:

- Goslate
- Textblob
- KMeans

For the map visualisation, Folium library was used with a tile layer of *'cartodbpositron'*. The nodes were marked within a boundary of *lat_min, lat_max & lon_min, lon_max* (calculated from the dataset coordinates).

EXPERIMENT:

We prepared a python code model to accept the csv data file we will be working on. We experimented on datasets: "Guwahati.csv" containing tweet data from Guwahati region and "Silchar.csv" containing tweet data from Silchar region.

Our first challenge was to translate the texts written in local languages like Assamese, Bengali or Hindi or in some cases, a different language entirely. Not only this, the linguistic approach also posed challenges like Devanagari words written in English script and phonetic transliteration, among many others. Fortunately for us, most of the translation work could be entrusted upon the python dependency packages such as *langdetect* to detect the languages other than English, and

translate and *deep_translator* to translate those languages to English. We also experimented with *iNLTK*- a text processing python library for Indian languages. The output to this step was stored as a new csv file and in the end, to ensure that the translations were up to mark, manual verification was conducted upon first few and bottom few tweet data.

Secondly, we translated the emojis to English text using python library *emoji*. The purpose behind this was to consider the sentiment value of the emojis as well, for a more accurate decision making.

Finally, the dataset contained urls to social media posts which were crucial in calculating the exact sentiment of the particular user. It was observed that out of all social media urls present in the dataset, 99% of them were of instagram posts. Therefore, instagram API was used for fetching the captions of the posts using the Requests library. These captions were further demojized, translated & then concatenated with the final text.

The final csv generated, which will be used in the sentiment calculation, is stored in the 'final datasets' directory

ANALYSIS:

We obtained a fairly neat output for both the available datasets. Given the number of input data, the output nodes did not have to compete for space and visibility. All the nodes can be vividly observed on the map, and therefore be analysed without confusion.

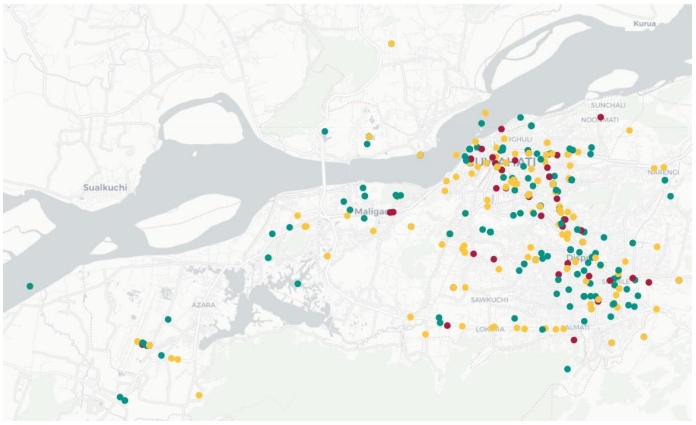

Figure 1. Representation of all nodes in Guwahati Dataset along with the 'Sentiment Score'

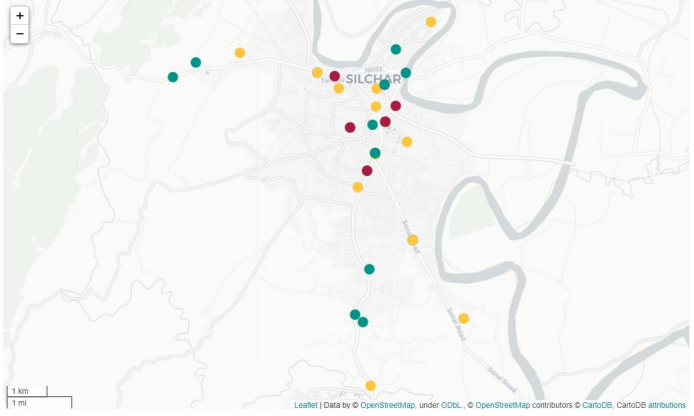

Figure 2. Representing all the nodes of the Silchar Dataset along with the 'Sentiment Score'

As we can see from **Figure 1 & 2**, the nodes with a positive Sentiment Score were marked green, nodes with negative Sentiment Score were marked red and finally nodes with a neutral score were marked yellow.

DATA CLUSTERING

Clustering is one of the most common exploratory data analysis techniques used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different.

KMeans Clustering. An iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group which in this case is the Sentiment Score (positive, negative & neutral).

Given an initial set of k means $m_1^{(1)},\ldots,m_k^{(1)}$ (see below), the algorithm proceeds by alternating between two steps:

Assignment step: Assign each observation to the cluster with the nearest mean: that with the least squared Euclidean distance (partitioning the clusters according to Voronoi Diagrams.

$$S_i^{(t)} = \{x_p : ||x_p - m_i^{(t)}||^2 \leq ||x_p - m_j^{(t)}||^2 \; \forall j,\ 1 \leq j \leq k\},$$

where x_p is assigned to exactly one $S^{(t)}$.

Update step: Recalculate centroids for observations assigned to each cluster.

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_j$$

Categorising the score in positive, negative and neutral, we represented the nodes as follows:

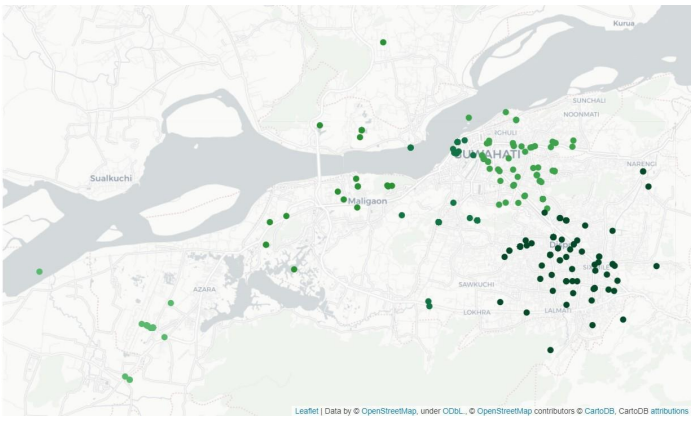

Figure 3. K Means clustering for nodes with positive Sentiment Score in Guwahati Dataset.

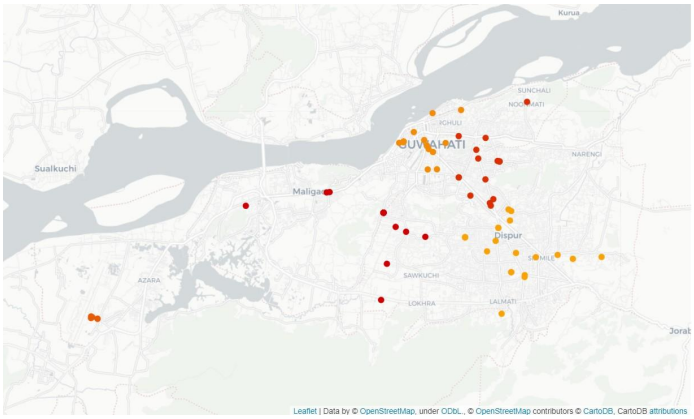

Figure 4. K Means clustering for nodes with negative Sentiment Score in Guwahati Dataset.

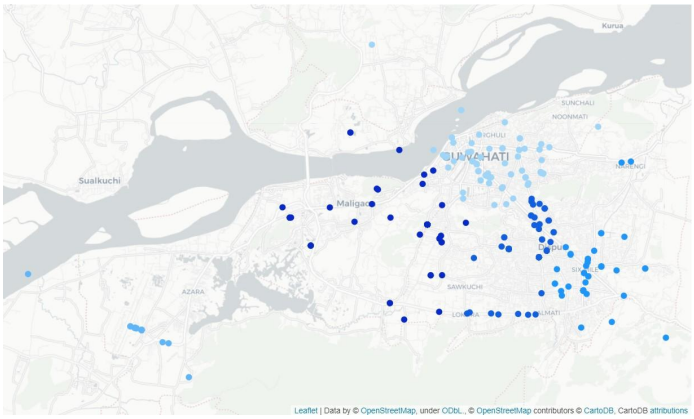

Figure 5. K Means clustering for nodes with neutral Sentiment Score in Guwahati Dataset.

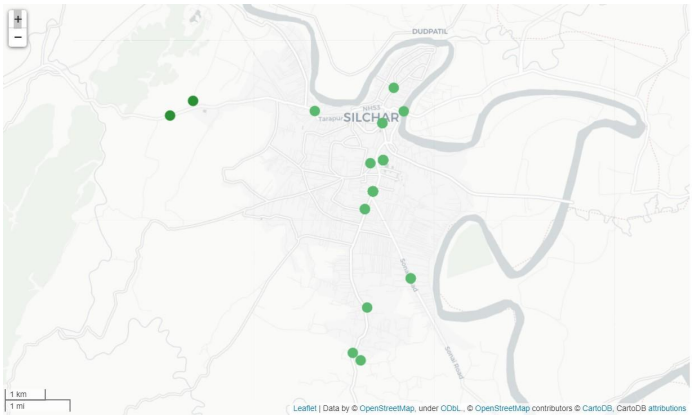

Figure 6. K Means clustering for nodes with positive Sentiment Score

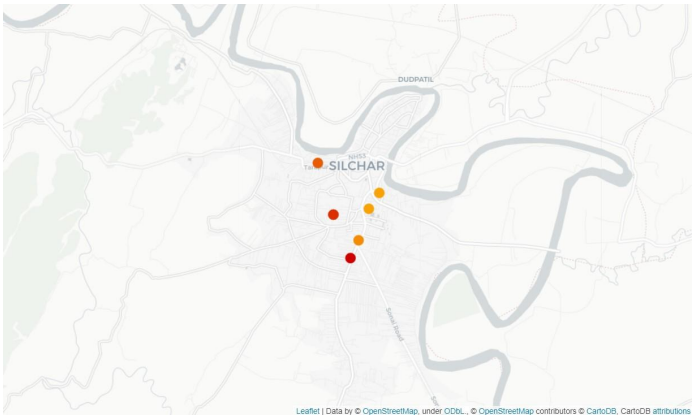

Figure 7. K Means clustering for nodes with negative Sentiment Score

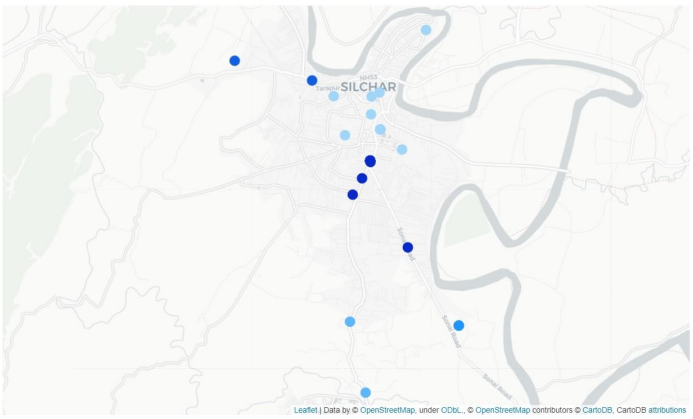

Figure 8. K Means clustering for nodes with neutral Sentiment Score in Silchar Dataset.

It must be noted that some of the obtained nodes were manually analysed for their placement accuracy on the basis of latitude and longitudinal values, as well as cross-checked for their sentiment accuracy by reading the tweet and giving a manual score.

Finally, word clouds were generated for both Guwahati and Silchar datasets. These figures were useful in depicting the thoughts and sentiments of the people by showing the bulk of the word used.

Figure 9. Word cloud for Guwahati

Figure 10. Word cloud for Silchar dataset.

CONCLUSION:

In this study, sentiment analysis was done on the pre-processed tweets taken from Guwahati & Silchar datasets. It was observed that **43.15% of the people living in Guwahati were happy**, while **15.02% were having a bad day** or undergoing some sad experience. Also, **41.83% appeared to be neutral** about their feelings or shared some information with neutral sentiment.

Likewise, for Silchar, it was observed that **41.4% of the people were happy**, while **13.44% had a negative sentiment** and **45.06% appeared to be neutral** in their tweets.

BIBLIOGRAPHY:

Related Works:

1. Amit Agrawal, Ritu Singh & Durga Toshniwal (2018) Geospatial sentiment analysis using twitter data for UK-EU referendum, Journal of Information and Optimization Sciences, 39:1, 303-317, DOI: 10.1080/02522667.2017.1374735 https://doi.org/10.1080/02522667.2017.1374735
2. Halldor Janetzko, Dominik J ´ ackle and Tobias Schreck (2013) Geo-Temporal Visual Analysis of Customer Feedback Data Based on Self-Organizing Sentiment Maps, International Journal On Advances in Intelligent Systems, International Academy, Research, and Industry Association (IARIA), 7(1 and 2):237--246, http://dx.doi.org/10.1177/1473871613481691

Reference Sites:

[1] Sentiment Analysis Definition, Wikipedia https://en.wikipedia.org/wiki/Sentiment_analysis

[2] What is Geospatial? MAPPS https://www.mapps.org/page/WhatisGeospatial#:~:text=In%20simple%20terms%2C%20geospatial%20information,glove%20compartment%20of%20your%20car

[3] Geospatial Data Analytics, Safegraph https://www.safegraph.com/guides/geospatial-data-analytics

[4] KMeans Clustering, Algorithm, Applications, Evaluation Methods, and Drawbacks

https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a