**Early Detection and Tracking of Disease Outbreaks using AI and Social Media Data**

**DATA REPORT**

**November 2024**

**TABLE OF CONTENT**

**Business Understanding……………………………………………….1**

**Data Understanding……………………………………………………..2**

**Data Preprocessing……………………………………………………...3**

**Modelling…………………………………………………………………..4**

**Model Evaluation………………………………………………………..5**

**Recommendations ……………………………………………………….6**

**Next Step……………………………………………………………………..7**

**1.Business Understanding**

**1.1Business overview**

This AI-driven detection system leverages social media data to track outbreaks of influenza and tuberculosis, two significant public health threats with widespread respiratory impacts. By analyzing real-time posts for early signs of disease clusters, the system enables WHO and health agencies to identify potential outbreaks quickly. This proactive approach aims to reduce disease spread, improve response times, and ultimately lower morbidity and mortality rates

**1.2 Problem Statement**

Early detection of influenza and tuberculosis outbreaks is critical, but official reporting often lags behind the initial spread. Social media data could provide an early warning system, but requires AI-powered analysis to identify and track disease trends before they become full-blown epidemics.

**1.3 Objectives**

**1.3.1.Main Objective**

Early detection of potential influenza and tuberculosis outbreaks using real-time twitter data.

**1.3.2 Specific Objectives**

1. Track the spread patterns of these diseases by monitoring symptom-related keywords
2. Develop predictive models to forecast the trajectory of potential outbreaks.
3. Provide early alerts to public health organizations and government agencies to enable faster response and intervention.

**1.4 Proposed Solutions**

1. Collect real-time social media data from Twitter using Twibot: and keyword-based filtering.
2. Apply natural language processing (NLP) techniques to detect mentions of disease symptoms, concerns, and outbreak-related keywords.
3. Conduct sentiment analysis to identify posts indicating fear, panic or growing anxiety around potential outbreaks.
4. Leverage machine learning models like SVMs and neural networks to classify social posts as related to disease outbreaks or not.
5. Map the geospatial and temporal data to visualize disease spread patterns and high-risk clusters.
6. Develop predictive models to forecast outbreak trajectories and build an automated alert system for public health authorities.

**1.5 Metric of success**

The accuracy is 90%, and the recall is 90%

**1.6 Stakeholders**

1. World Health Organization (WHO) Global health agency responsible for coordinating pandemic preparedness and response
2. National/Regional Public Health Organizations Disease control centers, epidemiology departments in countries/regions
3. Emergency Response Agencies Disaster management authorities, emergency medical teams
4. Government Policymakers Health ministers, legislators responsible for public health policies
5. Healthcare Providers Hospitals, clinics, and other medical facilities that need early warning
6. Public Health Researchers and Epidemiologists Academics and analysts studying disease trends and mitigation strategies
7. The General Public Citizen stakeholders who benefit from faster outbreak response and containment

**2.Data understanding**

Data Sources The data originates **twitter**.The data was scraped using **Twibot** Below is the description of the column names.The dataset has 11600 rows and 20 columns .

|  |  |
| --- | --- |
| **Columns** | **Description** |
| id | Unique identifier for each tweet entry in the dataset. |
| tweetText | Text content of the tweet. |
| tweetURL | URL link to the original tweet on Twitter. |
| type | Type of social media post, in this case all are tweets. |
| tweetAuthor | Name of the person or organization that posted the tweet. |
| Handle | Twitter handle or username of the tweet author. |
| geo | Geographical location associated with the tweet, if available. |
| mentions | Other Twitter users mentioned in the tweet. |
| hashtags | Hashtags used in the tweet. |
| replyCount | Number of replies to the tweet. |
| quotecount | Number of times the tweet was quoted by other tweets. |
| retweetCount | Number of times the tweet was retweeted. |
| likeCount | Number of likes the tweet received. |
| views | Number of times the tweet was viewed. |
| bookmarkcount | Number of times the tweet was bookmarked by users. |
| createdAt | Date and time when the tweet was created. |
| allMediaURL | URL to any media content (images, videos) associated with the tweet. |
| videoURL | URL specifically for video content in the tweet, if available. |
| Source\_file | Name of the file from which this tweet data entry originated. |
| Combined\_timestamp | Timestamp combining both the date and time of data processing or entry. |

**3. Data Preprocessing**

**3.1 Data Cleaning**

During the data cleaning process, we will remove white spaces from column names, check for missing values in each column, and look for duplicate entries.

SHAPE

Records in dataset: 11600 with 20 columns.

**UNIQUE VALUES**

* Column id has 11600 unique values
* Column tweetText has 6479 unique values
* Column tweetURL has 6501 unique values
* Column type has 1 unique values

**Top unique values in type include:**

* Tweet
* Column tweetAuthor has 5070 unique values
* Column handle has 5140 unique values
* Column geo has 2420 unique values
* Column mentions has 2588 unique values
* Column hashtags has 1204 unique values
* Column replyCount has 56 unique values
* Column quoteCount has 21 unique values
* Column retweetCount has 79 unique values
* Column likeCount has 158 unique values
* Column views has 1034 unique values
* Column bookmarkCount has 43 unique values
* Column createdAt has 6432 unique values
* Column allMediaURL has 1754 unique values
* Column videoURL has 284 unique values
* Column source\_file has 2 unique values

**MISSING VALUES**

* Column id has 0 missing values.
* Column tweetText has 0 missing values.
* Column tweetURL has 0 missing values.
* Column type has 0 missing values.
* Column tweetAuthor has 0 missing values.
* Column handle has 0 missing values.
* Column geo has 3772 missing values.
* Column mentions has 5725 missing values.
* Column hashtags has 9247 missing values.
* Column replyCount has 0 missing values.
* Column quoteCount has 0 missing values.
* Column retweetCount has 0 missing values.
* Column likeCount has 0 missing values.
* Column views has 0 missing values.
* Column bookmarkCount has 0 missing values.
* Column createdAt has 0 missing values.
* Column allMediaURL has 8831 missing values.
* Column videoURL has 11105 missing values.
* Column source\_file has 0 missing values.
* Column combined\_timestamp has 0 missing values.

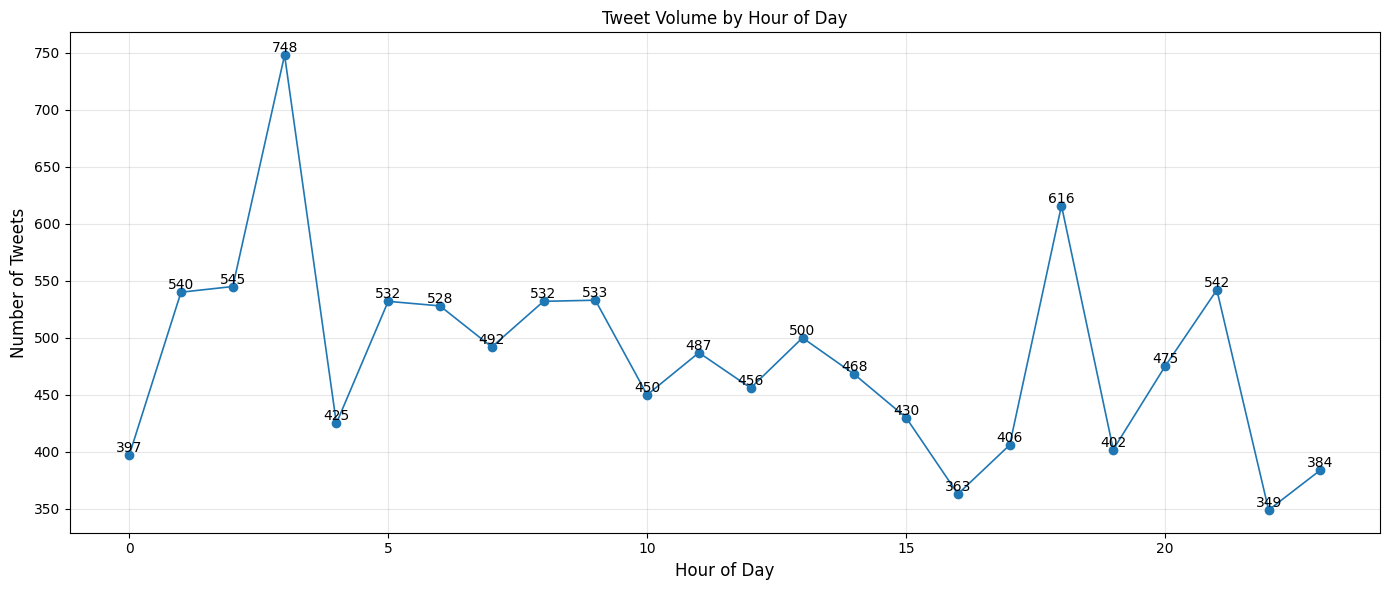
**DUPLICATE VALUES**

----------------

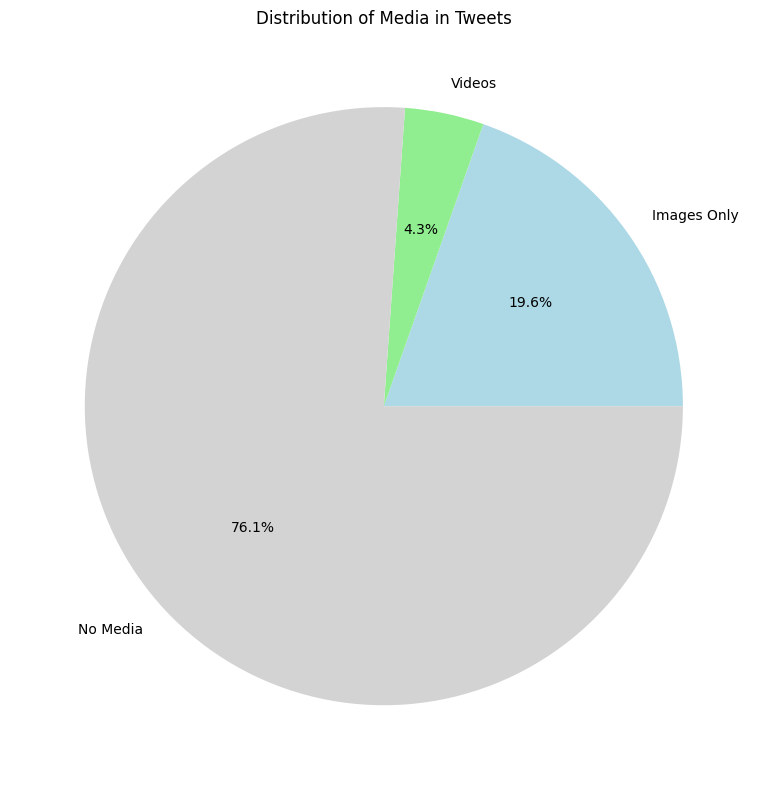
The dataset has 0 duplicated records.

**3.1 Data Visualization**

*****Time series analysis of tweet volume:(A lineplot)*****

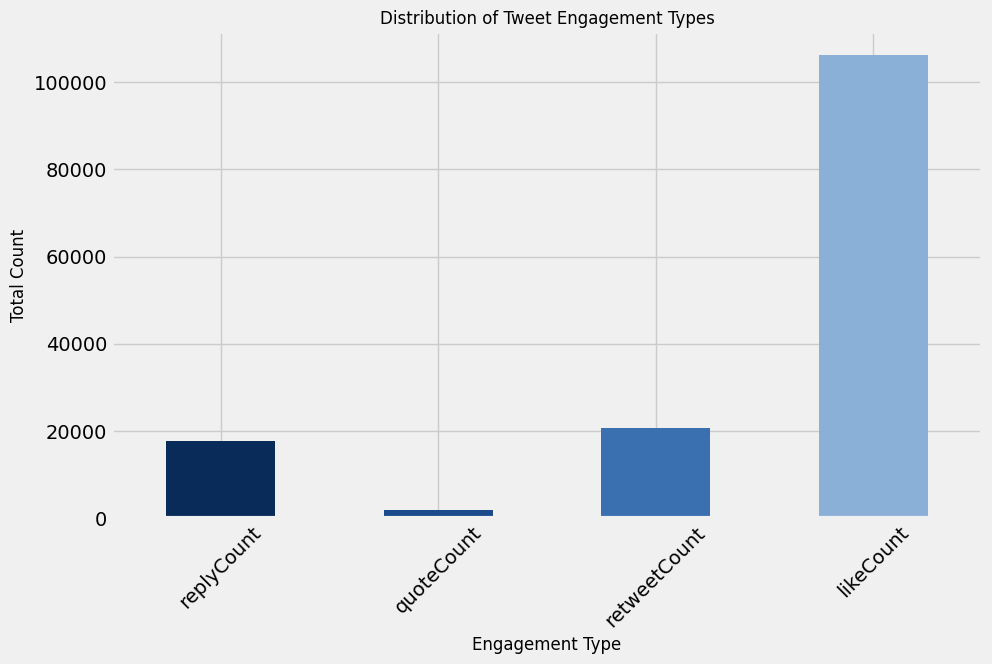


* Users are most active in the early morning (possibly automated posts or scheduled content)
* There's another surge of activity during evening hours when people are likely done with work
* The lowest engagement is in the very early morning hours when most people are sleeping
* There's consistent moderate activity during typical working hours



* No Media (82.2%): The vast majority of tweets are text-only, containing no images or videos. This shows that plain text remains the dominant form of communication on the platform.
* Images Only (14.2%): About one in seven tweets contains at least one image, making it the most common form of media content.
* Videos (3.7%): A small portion of tweets include video content, representing the least common media type.

*****Tweet Engagement Distribution*****



****Likes (likeCount):****

* By far the most common form of engagement with approximately 77,000 total likes
* Represents the easiest and most casual way to interact with a tweet

****Replies (replyCount):****

* Second most common engagement type with about 14,000 replies Represents a more active form of engagement as it requires users to write a response

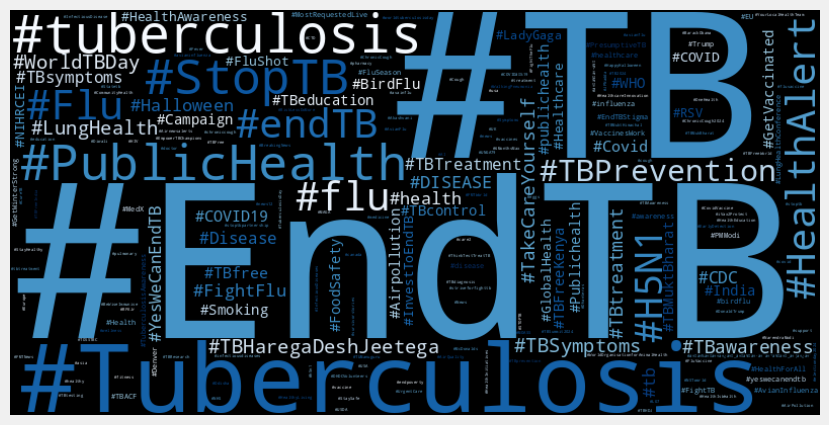
****Retweets (retweetCount):****

* Third most common with roughly 13,000 retweets Shows how often content is shared with users' followers

****Quotes (quoteCount):****

Least common with only about 2,000 quote tweets Represents the most involved form of engagement as it requires both sharing and adding commentary

*****Top Hashtags Word Cloud*****



## 4. Modelling

*****Models used:*****

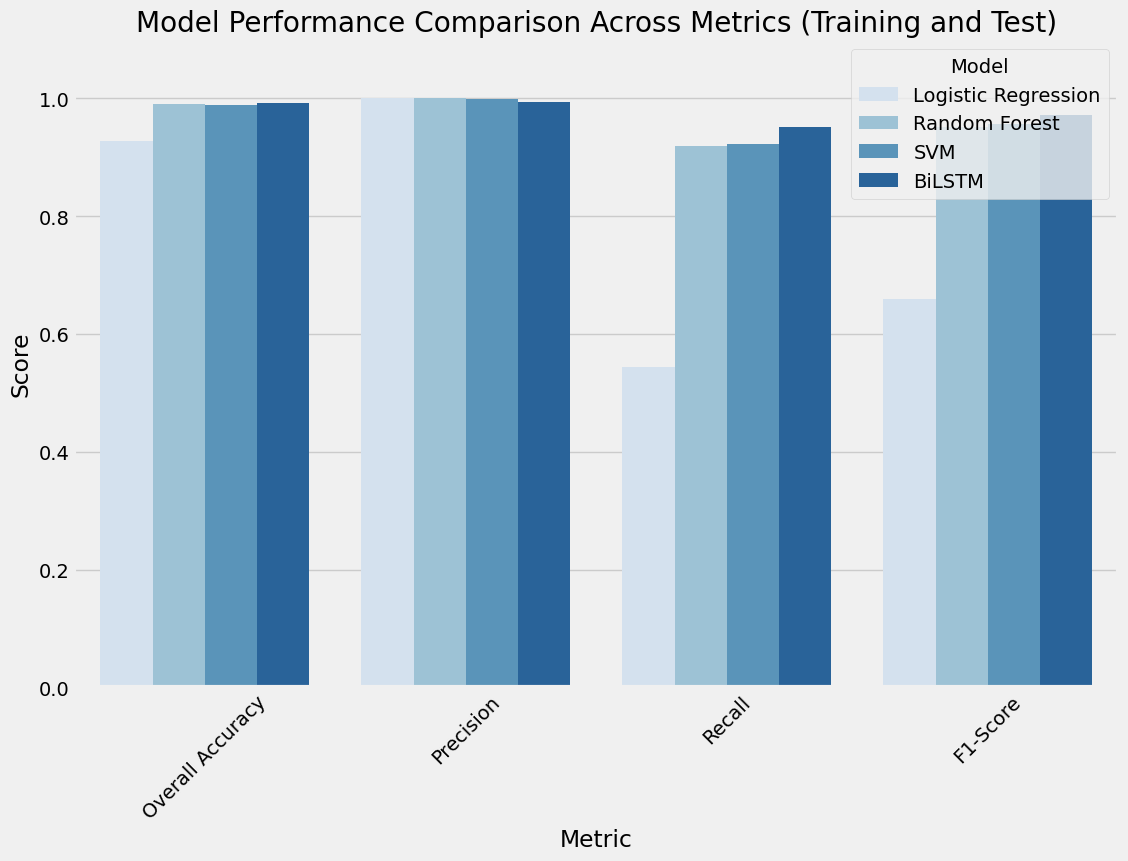
* Logistic Regression
* Random Forest
* SVM (Support Vector Machine)
* BiLSTM (Bidirectional Long Short-Term Memory)

*****1. Logistic Regression Model*****

Logistic Regression-based symptom detection model using TF-IDF features. It defines a SymptomDetectionLogistic class that:

* Preprocesses text data using TF-IDF vectorization
* Trains a multi-output Logistic Regression classifier
* Evaluates the model on training and test sets
* Saves the trained model to a pickle file

**Model Performance**



**5. Conclusion**

****Recall:****

* The Random Forest model achieved the highest recall score on the training set (1.0000) and a strong recall on the test set (0.8368), though it’s slightly lower than BiLSTM on the test set (0.9150).
* BiLSTM also performed well in recall, with a score of 0.9886 on the training set and 0.9150 on the test set. Conclusion: BiLSTM had the best generalization in terms of recall on the test set.

****Accuracy:****

* The Random Forest model achieved perfect accuracy on the training set (1.0000) and an impressive 0.9823 on the test set.
* The BiLSTM model also performed exceptionally, with 0.9984 accuracy on the training set and 0.9858 on the test set. Conclusion: BiLSTM slightly outperformed Random Forest on the test set in terms of accuracy.

****Overall Best Model:****

* Considering both recall and accuracy on the test set, the BiLSTM model emerges as the best-performing model due to its strong test set performance, which indicates better generalization to unseen data.

**6. Recommendations**

1. Track model performance metrics over time
2. Set up alerts for performance degradation
3. Regularly tune and retrain your machine learning models to maintain high accuracy, especially as new data trends emerge.

**7.Next Step**

1. Incorporate Related Social Media Platforms: Expand data collection to other platforms (like Facebook or Instagram) if applicable. This broadens the dataset and captures a wider public sentiment.

2. Establish Partnerships with Health Agencies: Partner with health organizations that could benefit from timely disease information, enhancing the project's real-world impact.

3. Develop Clear Data Retention Policies: Define how long data will be stored, particularly sensitive information like location, to ensure ethical data handling.

4. Expand Symptom Detection: Broaden the system’s capability to detect symptoms of other infectious diseases, enhancing its value for monitoring various health conditions.

5. Improve Detection Accuracy: Use additional training data and advanced machine learning techniques to improve the accuracy of detecting tuberculosis symptoms.

6. Real-Time Alerts: Set up real-time alert systems to notify health authorities instantly when potential cases are detected, enabling quicker responses.

7. Multilingual Support: Enable the system to analyze posts in multiple languages, allowing for broader monitoring across regions with different languages.

8. Collaboration with Health Agencies: Partner with WHO and public health agencies to align the system with public health standards and ensure it meets practical needs for outbreak management.