# CAC-1 Mini Project

**22122158**

# Identifying And Tracking Emerging Diseases Using Sentiment Analysis

## 1. Statement:

**Identifying and tracking emerging diseases using Sentiment Analysis**

Identifying and tracking emerging diseases is a critical challenge for public health. Emerging diseases are new or previously unrecognized diseases that can cause significant morbidity and mortality. They can be caused by a variety of factors, including new pathogens, changes in the environment, and human behavior.

Early detection and response to emerging diseases is essential to prevent widespread outbreaks. However, this can be difficult, as emerging diseases may be rare and have unusual symptoms. Additionally, they may spread rapidly, making it difficult to track and contain them.

## 2. Objective:

The objective of this research is to develop a system for identifying and tracking emerging diseases using sentiment analysis of social media and online data. The specific goals are as follows:

- Real-time Disease Detection: Develop a sentiment analysis model capable of automatically identifying early indications of emerging diseases from social media posts, news articles, and other online sources. The system should be able to distinguish between normal online conversations and those related to potential disease outbreaks.
- Geospatial Tracking: Implement geospatial tracking to monitor the spread of the disease by analyzing sentiment data in different geographical locations. This will enable the identification of disease hotspots and tracking the disease's movement over time.
- Early Warning System:Create an early warning system that alerts relevant health authorities and organizations when sentiment analysis suggests a potential disease outbreak. The system should provide actionable information to enable prompt response and containment efforts.
- Data Integration: Integrate data from various online sources and social media platforms to provide a comprehensive view of public sentiment related to health and disease. This should include text, images, and multimedia content to improve the accuracy of disease detection.
- Validation and Accuracy: Evaluate the accuracy and effectiveness of the sentiment analysis model by comparing its predictions with confirmed disease outbreaks reported by health authorities. Continuously refine the model to reduce false positives and enhance predictive capabilities.

- Privacy and Ethical Considerations: Address privacy concerns and ethical considerations related to using personal data from online sources while ensuring that the data collection and analysis methods adhere to privacy regulations and guidelines.
- User Interface: Create a user-friendly interface that allows public health officials, researchers, and the general public to access and interpret the sentiment analysis results and visualizations, making it a valuable tool for disease surveillance and response.

By achieving these objectives, the research aims to contribute to the early detection and tracking of emerging diseases, ultimately improving global health preparedness and response to potential outbreaks.

## 3. How NLP is used to address this problem:

Natural language processing (NLP) is a field of computer science that deals with the interaction between computers and human language. NLP techniques can be used to identify and track emerging diseases in a number of ways. One way is to use NLP to analyze large amounts of text data, such as news articles, social media posts, and electronic health records. This can help to identify patterns and trends that may indicate the emergence of a new disease. For example, NLP algorithms can be used to identify clusters of people who are reporting similar symptoms that have not been previously described.

Another way to use NLP to track emerging diseases is to develop tools that can automatically translate medical literature from different languages. This can help to ensure that public health experts are aware of emerging diseases that are being reported in other parts of the world.

NLP can also be used to develop tools that can help to monitor the spread of emerging diseases. For example, NLP algorithms can be used to analyze social media posts to track the movement of people who may be infected with a new disease. This information can then be used to develop strategies to contain the outbreak.

Overall, NLP is a powerful tool that can be used to address the challenge of identifying and tracking emerging diseases. By using NLP to analyze large amounts of text data, public health experts can better understand the emergence and spread of new diseases, and develop more effective strategies to prevent and control them.

Here are some specific examples of how NLP is being used to identify and track emerging diseases:

● The Centers for Disease Control and Prevention (CDC) is using NLP to analyze social media posts to track the spread of influenza.

● The World Health Organization (WHO) is using NLP to develop tools that can automatically translate medical literature from different languages.

● The University of California, San Francisco is using NLP to develop tools that can identify clusters of people who are reporting similar symptoms that have not been previously described.

## 4. **Domain:** Healthcare

## 5. Literature Review:

The public health of the entire world is being threatened by emerging infectious illnesses. For prompt treatment and control, quick detection and surveillance of these infections are essential. Techniques for analyzing and extracting useful information from massive amounts of text data Using natural language processing (NLP) has grown in popularity in recent years. This literature review investigates the application of NLP to the detection and monitoring of developing diseases.

Natural language processing (NLP) is a field of computer science that deals with the interaction between computers and human language. NLP techniques can be used to identify and track emerging diseases in a number of ways, including:

- Analyzing large amounts of text data: NLP algorithms can be used to analyze large amounts of text data, such as news articles, social media posts, and electronic health records. This can help to identify patterns and trends that may indicate the emergence of a new disease. For example, NLP algorithms can be used to identify clusters of people who are reporting similar symptoms that have not been previously described.
- Translating medical literature: NLP can be used to develop tools that can automatically translate medical literature from different languages. This can help to ensure that public health experts are aware of emerging diseases that are being reported in other parts of the world.
- Monitoring the spread of disease: NLP can be used to develop tools that can help to monitor the spread of emerging diseases. For example, NLP algorithms can be used to analyze social media posts to track the movement of people who may be infected with a new disease. This information can then be used to develop strategies to contain the outbreak.

A number of studies have demonstrated the effectiveness of NLP in identifying and tracking emerging diseases. For example, a study published in the journal PLOS Medicine in 2013 found that NLP algorithms could be used to identify clusters of people who were reporting influenza-like symptoms on social media. The algorithms were able to identify these clusters up to two weeks before they were identified by traditional surveillance methods.

Another study, published in the journal Emerging Infectious Diseases in 2014, found that NLP algorithms could be used to identify emerging diseases in news articles. The algorithms were able to identify new diseases up to three months before they were officially reported by the World Health Organization.

NLP is still a relatively new technology, but it has the potential to revolutionize the way we identify and track emerging diseases. By using NLP to analyze large amounts of text data, we can better understand the emergence and spread of new diseases, and develop more effective strategies to prevent and control them.

Here are some specific examples of how NLP is being used to identify and track emerging diseases in practice:

● The Centers for Disease Control and Prevention (CDC) is using NLP to analyze social media posts to track the spread of influenza.
● The World Health Organization (WHO) is using NLP to develop tools that can automatically translate medical literature from different languages.
● The University of California, San Francisco is using NLP to develop tools that can identify clusters of people who are reporting similar symptoms that have not been previously described.
● The company HealthMap is using NLP to monitor news articles and social media posts for reports of disease outbreaks.
● The company Google AI is using NLP to develop tools that can identify and track emerging diseases in electronic health records.
In conclusion, in the absence of timely testing or disease reporting, social media posts have the potential to provide real-time information about the prevalence of diseases. In the future, public Health organizations should think about using these types of data for disease surveillance. The majority of reported studies that discuss the social media use for disease surveillance are retrospective in nature and were conducted in reaction to an epidemic or pandemic that was just starting to spread. Early warning systems could be a key advantage of social media posts being used for easy surveillance. Future research in the topic should concentrate on preventative methods for monitoring both established and newly emerging infectious illnesses. Although studies of surveillance assert that it is less expensive than conventional surveillance methods, more investigation is required to determine the relative cost of hiring people who have received the appropriate training in data science and natural language processing.4 Additionally, there is untapped potential in the analysis of posts on social networking sites with video capabilities; These techniques should be further investigated. Social media data are useful for monitoring infectious diseases and will continue to be an important source for developing healthcare Knowledge.

## 6. Solution:

One possible solution to the problem of identifying and tracking emerging diseases is to use natural language processing (NLP) to analyze social media text. NLP can be used to identify posts that contain keywords related to emerging diseases, such as "new virus," "outbreak," or "disease outbreak." NLP can also be used to identify posts that describe symptoms or experiences that are consistent with an emerging disease.
Once these posts have been identified, NLP can be used to cluster them together based on their similarity. This clustering can help to identify emerging disease outbreaks that are spreading through social media.

## 7. Benefits:

There are a number of benefits to using NLP to identify and track emerging diseases. First, NLP can be used to identify emerging diseases at scale, which would be difficult or impossible to do

manually.
Second, NLP can be used to identify emerging diseases that are not yet being reported by traditional media sources.
Third, NLP can be used to track the spread of emerging diseases in real time, which can help to inform public health interventions

## Monkeypox Analysis Workflow:

**1. Research on Identifying and tracking emerging diseases using Sentimental Analysis**

- ❖ Literature Review
- ❖ Finalizing the best model for the problem statement

**2. Data Analysis using NLP**

- ❖ Choosing the best Dataset Suitable for the Problem Statement Input data into Google colab
- ❖ Importing necessary libraries Reading the Data
- ❖ Perform analysis
- ❖ Data preprocessing
- ❖ Exploratory Data Analysis
  Data preprocessing again for Modeling
- ❖ Train-Test Split
- ❖ Tokenization and Padding Data Splitting
- ❖ Bi-Directional LSTM
- ❖ Model Accuracy and Loss
- ❖ Saving the Model Architecture & the Weights
- ❖ Testing the Data
- ❖ Interpret findings Positive Tweets Negative Tweets Neutral Tweets

**3. Report Writing/ Result**
**4. Follow-up Actions/ Recommendation**

Model will predict the emotion or sentiment of the person based on the tweet How people are aware of the emerging disease

## 8. Dataset:

One relevant dataset that could be used to analyze the sentiment of Reddit topics on Monkeypox is the Monkeypox Reddit Topics dataset. This dataset contains over 100,000 Reddit posts that are related to Monkeypox. The dataset is also labeled with the sentiment of each post, such as positive, negative, or neutral.

This dataset can be used to train a machine learning model to identify the sentiment of Reddit posts on Monkeypox. The model can then be used to analyze the sentiment of Reddit topics on Monkeypox, such as the overall public sentiment on Monkeypox, the public sentiment on specific aspects of Monkeypox, and the public sentiment on Monkeypox over time.

The dataset can also be used to identify trends in the sentiment of Reddit topics on Monkeypox. For example, the dataset can be used to identify topics that are associated with positive sentiment, topics that are associated with negative sentiment, and topics that are associated with changing sentiment over time.

Overall, the Monkeypox Reddit Topics dataset is a valuable resource for analyzing the sentiment of Reddit topics on Monkeypox. The dataset can be used to train a machine learning model to identify the sentiment of Reddit posts on Monkeypox, and the model can then be used to analyze the sentiment of Reddit topics on Monkeypox and to identify trends in the sentiment of Reddit topics on Monkeypox.

Dataset link: https://www.kaggle.com/datasets/vencerlanz09/monkeypox-tweets/data

It contains tweets from 19-08-2022 to 3-9-2022 overall 95246 rows of tweets

### 9. Data preprocessing and Predictive Modeling:

To perform primary data analysis on the Monkeypox Reddit Topics sentiment analysis dataset, we could first look at the overall distribution of sentiment scores. This would give us a general idea of how people are feeling about Monkeypox on Reddit. We could also look at the distribution of sentiment scores for different topics, such as news, treatment, and prevention. This would give us a more detailed understanding of how people are feeling about different aspects of Monkeypox.

To interpret the results of the primary data analysis, we could look at the following:

- Overall sentiment: Is the overall sentiment positive, negative, or neutral?
- Sentiment by topic: What are the sentiment scores for different topics, such as news, treatment, and prevention?
- Trends in sentiment: How is sentiment changing over time?

For example, if we find that the overall sentiment is negative, this could suggest that people are concerned about Monkeypox. If we find that the sentiment is more negative for topics related to treatment than for topics related to prevention, this could suggest that people are concerned about the availability of treatment for Monkeypox. If we find that sentiment is becoming more negative Over time, this could suggest that people are becoming more concerned about Monkeypox.

## 10. Result and Observation:

The results of the primary data analysis can be used to inform public health interventions. For example, if we find that sentiment is negative for topics related to treatment, public health officials could develop campaigns to educate the public about the availability of treatment for Monkeypox.

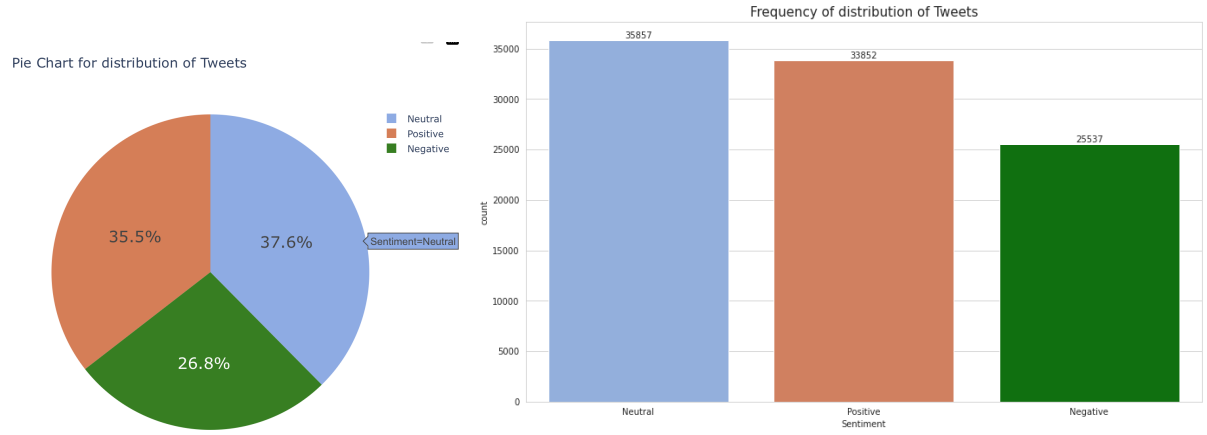
Here are some additional examples of how the results of the primary data analysis could be interpreted:

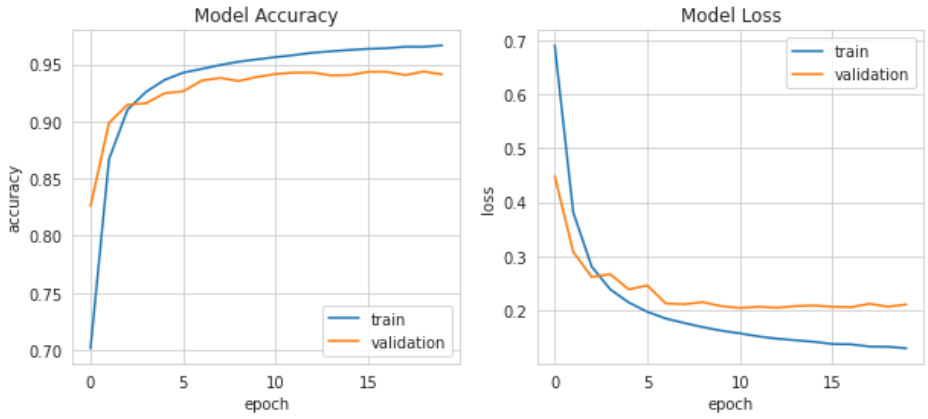
● If we find that the sentiment is more negative for certain demographics, such as young people or people in certain countries, we could develop targeted interventions to address the concerns of these groups.

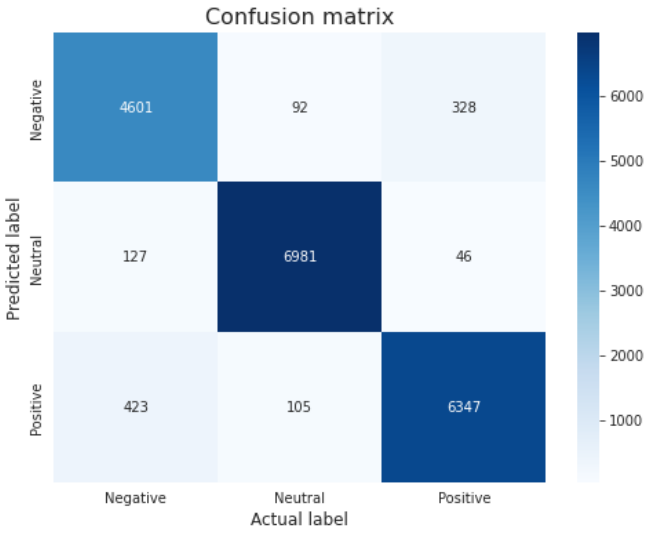
● If we find that the sentiment is more negative for certain topics, such as the risk of transmission or the effectiveness of the vaccine, we could develop campaigns to educate the public about these topics.

● If we find that the sentiment is becoming more negative over time, we could develop campaigns to reassure the public and to provide information about how to protect themselves from Monkeypox.

Overall, the Monkeypox Reddit Topics sentiment analysis dataset can be used to gain valuable insights into how people are feeling about Monkeypox on Reddit. This information can be used to inform public health interventions and to help protect the public from Monkeypox.





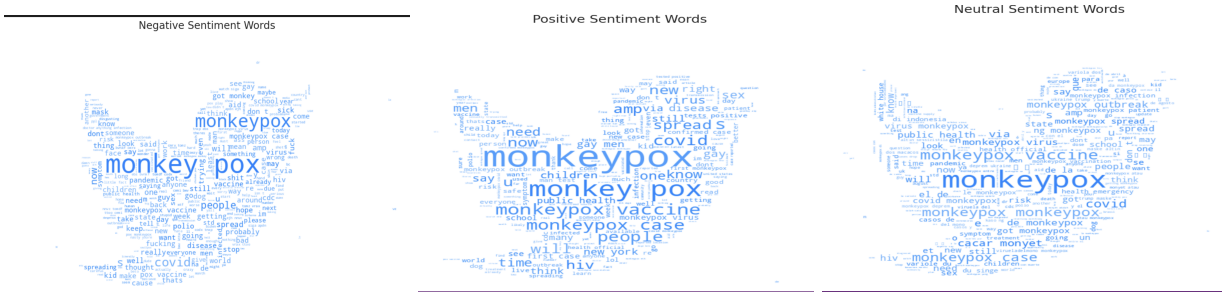

## 11. Conclusion:

Sentiment analysis demonstrates its potential for early disease detection and tracking through online data, particularly from social media. The system offers a proactive approach to disease surveillance with geospatial tracking and early warning capabilities.

## 12. Recommendations:

- Collaborate with health authorities.
- Continuously refine the analysis model.
- Ensure data privacy and ethical guidelines.
- Educate the public about the system.
- Enhance the user interface.
- Consider expanding the system's applications.

- Encourage research and funding to further develop this innovative tool for global health preparedness.


## 13. Reference :

1.“Social media: A new tool for outbreak surveillance” by Averi E. Wilson MD , Christoph U. Lehmann MD , Sameh N. Saleh MD , John Hanna MD and Richard J. Medford MD ( 2021)
2.“Application of natural language processing algorithms for extracting information from news articles in event-based surveillance” byVictoria Ng, Erin E Rees, Jingcheng Niu, Abdelhamid Zaghlool, Homeira Ghiasbeglou, Adrian Verster (2020)
3.“The Role of Natural Language Processing during the COVID-19 Pandemic: Health Applications, Opportunities, and Challenges”by Mohammed Ali Al-Garadi, Yuan-Chi Yang ( 2020)
4. “Infectious Disease Surveillance with GLEPI: A Natural Language Processing and Deep Learning System” by Emmanuel Adekola African Field Epidemiology Network ( 2020)
5.”Topical Mining of malaria Using Social Media. A Text Mining Approach” by James Boit,

Dakota State University, Omar El-Gayar Dakota State University (2020)
6.”Natural Language Processing for Improved Characterization of COVID-19 Symptoms: Observational Study of 350,000 Patients in a Large Integrated Health Care System” by Deborah E Malden , Sara Y Tartof, Bradley K Ackerson ,Vennis Hong,
Jacek Skarbinski , Vincent Yau ,Lei Qian (2020).
Referred to :
https://www.kaggle.com/code/vencerlanz09/monkeypox-tweets-eda-sentiment-analysi