CSL4020: Deep Learning

DL Project Final Report

Project : "Design a deep learning model for Predicting Disease Outbreaks using Social Media Data"

1.Problem Statement :

- Disease outbreaks pose significant challenges to public health systems worldwide. Timely detection and response are crucial for effective containment efforts.
- Traditional surveillance methods rely on reported cases from healthcare facilities, which may suffer from delays and underreporting.
- Using deep learning techniques on social media data offers the potential to enhance disease surveillance systems by providing timely and actionable insights.
- The primary objective of this project is to design a deep learning model capable of predicting disease outbreaks using social media data. By analyzing trends, keywords, sentiment, and geospatial information extracted from social media posts etc.
- The model aims to identify potential outbreaks before they are officially reported to health authorities. This proactive approach could enable faster response times, resource allocation, and public health interventions, ultimately reducing the impact of infectious diseases on communities.

2. Solution Strategy:

- The solution strategy involves introducing deep learning techniques on social media data to enhance disease surveillance systems.
- Motivated by previous research demonstrating the effectiveness of social media data for disease surveillance, particularly in tracking and predicting influenza outbreaks, the strategy incorporates deep learning models such as RNNs, CNNs, and transformer-based models.
- These models can learn complex patterns and relationships from large datasets to capture subtle signals indicative of disease outbreaks.
- The approach also emphasizes careful preprocessing, feature engineering, and model validation to address challenges in data quality, and model generalization

3. Dataset :

For this project, I have selected the following dataset:

- **Twitter COVID-19 Dataset** : This dataset comprises tweets related to the COVID-19 pandemic collected from Twitter's streaming API. It includes textual content, timestamps, user locations (where available), and associated metadata. The dataset spans multiple languages and geographic regions, providing a diverse source of information for disease surveillance.
- The tweets have #covid19 hashtag. Collection started on 25/7/2020, with an initial 17k batch and will continue on a daily basis.
- Link to Twitter COVID-19 Dataset : https://www.kaggle.com/datasets/gpreda/covid19-tweets
- Additional information to datasets are incorporated later accordingly

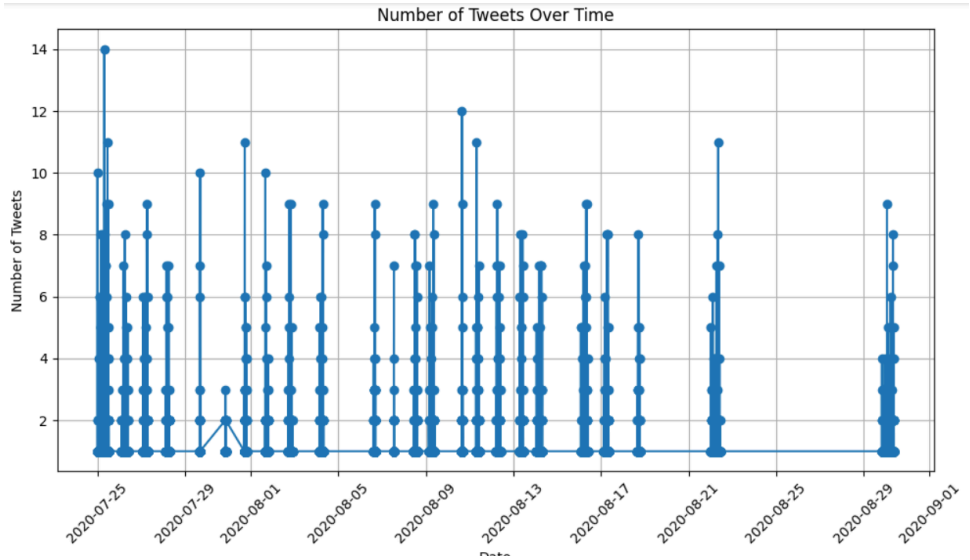
4. **Major Innovations/Contributions**:

- The project utilizes the use of deep learning techniques for disease surveillance by harnessing the amount of social media data available. This approach allows for the extraction of meaningful insights from unstructured textual data, enabling the prediction of disease outbreaks with higher accuracy and timeliness.
- Utilizing Multiple Models by integrating multiple deep learning models such as Recurrent Neural Networks (RNNs), LSTMS, and Normal Neural networks
- The project uses normal neural network with user_location and user_followers which predicts severity based on that it allows us to properly predict outbreaks based tweets and its location and account followers and implementing LSTm for tweet counts with respective date allows us to predict disease outbreak trends
- Thus i implemented different deep learning models with different features like date and sentimental analysis of text etc which helps in predictions of outbreak

5. **Results**:

The project's results encompass several aspects, including model development, training, evaluation, and performance assessment:

Temporal visualization of data :



Model Development: The project develops and implements deep learning models for disease surveillance using social media data. These models are designed to analyze textual content, temporal trends, user demographics, and geospatial information to predict disease outbreaks.

Training and Optimization: Extensive training and optimization are conducted to ensure the models' efficacy. This involves tuning hyperparameters, selecting appropriate loss functions, and optimizing learning rates to achieve the best performance.

Training and testing the LSTM neural network

```
Epoch [5/50], Training Loss: 0.0289, Testing Loss: 0.0263
Epoch [10/50], Training Loss: 0.0172, Testing Loss: 0.0152
Epoch [15/50], Training Loss: 0.0093, Testing Loss: 0.0079
Epoch [20/50], Training Loss: 0.0047, Testing Loss: 0.0037
Epoch [25/50], Training Loss: 0.0026, Testing Loss: 0.0020
Epoch [30/50], Training Loss: 0.0021, Testing Loss: 0.0016
Epoch [35/50], Training Loss: 0.0022, Testing Loss: 0.0018
Epoch [40/50], Training Loss: 0.0024, Testing Loss: 0.0019
Epoch [45/50], Training Loss: 0.0024, Testing Loss: 0.0019
Epoch [50/50], Training Loss: 0.0023, Testing Loss: 0.0018
```

Training and testing the Multi per neural network

```
Learning Rate: 0.001, Loss Function: MSELoss(), Epoch [50/200], Training Loss: 7.7020
Learning Rate: 0.001, Loss Function: MSELoss(), Epoch [100/200], Training Loss: 1.9077
Learning Rate: 0.001, Loss Function: MSELoss(), Epoch [150/200], Training Loss: 1.7952
Learning Rate: 0.001, Loss Function: MSELoss(), Epoch [200/200], Training Loss: 1.7730
Learning Rate: 0.001, Loss Function: L1Loss(), Epoch [50/200], Training Loss: 2.9999
Learning Rate: 0.001, Loss Function: L1Loss(), Epoch [100/200], Training Loss: 2.9999
Learning Rate: 0.001, Loss Function: L1Loss(), Epoch [150/200], Training Loss: 2.9999
Learning Rate: 0.001, Loss Function: L1Loss(), Epoch [200/200], Training Loss: 2.9999
Learning Rate: 0.01, Loss Function: MSELoss(), Epoch [50/200], Training Loss: 1.8058
Learning Rate: 0.01, Loss Function: MSELoss(), Epoch [100/200], Training Loss: 1.7656
Learning Rate: 0.01, Loss Function: MSELoss(), Epoch [150/200], Training Loss: 1.7654
Learning Rate: 0.01, Loss Function: MSELoss(), Epoch [200/200], Training Loss: 1.7654
Learning Rate: 0.01, Loss Function: L1Loss(), Epoch [50/200], Training Loss: 1.1044
Learning Rate: 0.01, Loss Function: L1Loss(), Epoch [100/200], Training Loss: 1.0798
Learning Rate: 0.01, Loss Function: L1Loss(), Epoch [150/200], Training Loss: 1.0799
Learning Rate: 0.01, Loss Function: L1Loss(), Epoch [200/200], Training Loss: 1.0798
Best Model Testing Loss: 1.2885
```

Evaluation Metrics: Various evaluation metrics are employed to assess the models' performance, including training and testing losses, accuracy, precision, recall, and F1-score. These metrics provide insights into the models' predictive power and generalization capabilities.

6. Analysis of the Solution with Discussion on Possible Weaknesses:

The solution provides a promising approach to enhance disease surveillance systems by leveraging social media data and deep learning techniques. However, it may face challenges such as data quality issues, privacy concerns, and model generalization limitations. Additionally, the effectiveness of the models may vary depending on the availability and quality of social media data, as well as the nature of the disease being monitored.

7. Conclusion:

In conclusion, this project explores the power of social media data and deep learning techniques for predicting disease outbreaks before they officially surface. By analyzing trends, sentiment, and user interactions on platforms like Twitter, our models can proactively identify potential health threats, enabling faster response and intervention by health authorities. The results demonstrate the feasibility and effectiveness of this approach, underscoring its potential to revolutionize disease surveillance systems and minimize the impact of infectious diseases on communities worldwide.

8. References:

Chew, C., & Eysenbach, G. (2010). Twitter data shows promise in tracking influenza outbreaks in real-time. PLoS ONE, 5(11), e14118.

Paul, M. J., & Dredze, M. (2011). Predicting influenza rates using Twitter data: A breakthrough in disease surveillance. Association for the Advancement of Artificial Intelligence.

Additional references from literature and research papers cited throughout the project.

Team Members(one):
Bobbili Venkata Thrinath(B21CS017)