INTERNSHIP REPORT

S Rethika

May 5, 2023 to Jun 30, 2023

OBJECTIVE

The objective of this project is to utilize unsupervised learning techniques to analyze COVID-19 death cases and develop an analytic model for estimating the number of days a patient will be admitted. By leveraging clustering algorithms and dimension reduction methods on a comprehensive dataset of COVID-19 fatalities obtained from https://stopcorona.tn.gov.in, the project aims to identify distinct patterns and correlations among various variables that influence the length of hospitalization such as the patients' age and the symptoms. The resulting analytic model will assist healthcare professionals in forecasting patient admissions and resource planning, enabling more efficient allocation of medical resources and better management of COVID-19 cases.

DATA EXTRACTION TO FINAL CLUSTERING

- Data was extracted from the PDFs ranging from the dates 01.01.2022 to 31.12.2022 using the PyPDF module, specifically focusing on extracting crucial information related to death cases and district-wise case details. This involved utilizing regular expressions and text extraction algorithms to extract relevant data points from the PDF files and make them into a CSV file. This enabled it to analyze and organize the death cases data as well as district-wise case details for further analysis and insights.
- The extracted data was then cleaned to eliminate any lexical errors and incorrect information. Each data point was reviewed to identify and rectify discrepancies, such as wrong values for dates and districts. Special attention was given to address common issues such as misspelled district names, inconsistent date formats and erroneous entries.
- After the cleaning process, the dataset was subjected to clustering analysis using KMeans, spectral clustering and DBSCAN algorithms. The clustering was solely based on the patient age as the feature. However the resulting clusters did not exhibit the desired outcomes due to the presence of diverse and varying data points.
- To enhance the clustering analysis, an additional feature was incorporated by encoding the district values using a label encoder. This enabled the inclusion of geographic information as a feature in the clustering process. Despite the inclusion

of an additional feature, the data did not exhibit distinct and well-separated clusters.

- In order to capture more meaningful patterns in the data, an alternative approach was taken by vectorizing the symptoms using TF-IDF (Term Frequency-Inverse Document Frequency). This allowed for the transformation of symptom text data into numerical feature vectors. However, despite this effort, an issue emerged where symptoms such as "Fever/Cough" and "Cough/Fever" were not clustered together due to the order in which the symptoms were presented.
- Before implementing the BERT model, I explored the functionality of neural networks by experimenting with both regression and classification models. This approach allowed me to gain a deeper understanding of how neural networks operate and how they can be applied to different types of problems. This knowledge served as a foundation for implementing more complex models, such as BERT to vectorize the symptoms.
- The BERT model was then implemented using the TensorFlow Hub module to preprocess the symptom data by tokenizing it and feeding it through the BERT model to obtain the pooled output. This approach addressed the limitations encountered with basic NLP techniques and TF-IDF. The tokenization process captured the context and relationships between the words, while the pooled output represented a fixed-length representation of the entire symptom sequence, capturing contextual information and semantic meaning.
- By incorporating the BERT model, the challenges related to order sensitivity of symptoms were overcome and improved the accuracy of the feature representation, resulting in enhanced clustering results and a deeper understanding of the underlying patterns in the symptom data.

CONCLUSION

In conclusion, the final clustering analysis yielded four distinct clusters, revealing significant variations in the average number of days admitted based on symptoms and age. This outcome demonstrates the effectiveness of incorporating the BERT model and TensorFlow Hub module to address order sensitivity and improve clustering results. The findings offer valuable insights into the relationship between symptoms, age, and the duration of hospitalization in COVID-19 cases, enabling more informed decision-making and resource allocation in healthcare management.