Areeba Umair

Devising Artificial Intelligence Tools for Complex Data

Tutor: Prof. Elio Masciari

Cycle: XXXVI

Year: Third

My background

- MS in Computer Science (MSCS) from Pakistan
- Currently PhD student of the ITEE program
- Ph.D. started on 1st November 2020
- UNINA Scholarship

Summary of study activities

Ad hoc PhD courses / schools

- Digital Forensics; methods, practices and tools
- Statistical data analysis for science and engineering research
- Software Defined Radio Applications for Radar and Localization Systems
- Ultra High Field Magnetic Resonance Imaging
- Safety Training Course
- Corso di Italiano livello A1
- AIRO PhD School 2021 and 5th AIRO Young Workshop
- 2021 Spring School on Transferable Skills
- Scuola Nazionale per Dottorandi “F. Gasparini”. XXIV Stage, Napoli

Courses attended borrowed from MSc curricula

- Data Visualization
- Hardware and Software Architectures for Big Data – Mod. B
- Big Data Analytics and Business Intelligence

Conferences / events attended

- 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), December 9, 2021, USA (Presented my paper online)
- 30th Euromicro International Conference on Parallel, Distributed and Network-Based Processing in Valladolid, Spain, March 9th - 11th, 2022 (Presented my paper Online)
- 30th Symposium on Advanced Database System - Tirrenia (Pisa), Italy - 19-22 June 2022 (Presented my paper in-person)
- ICIT 2022 International Conference on IT and Industrial Technologies, October 03-04, 2022

Summary of study activities

PhD Year	Courses	Seminars	Research	Tutoring / Supplementary Teaching	Total
First	20	5	35	0	60
Second	16	7	45	0	68
Third	02	0.4	60	0	62.4
Total	38	12.4	140	0	190.4
Expected	30-70	10-30	80-140	0-4.8	120-244.8

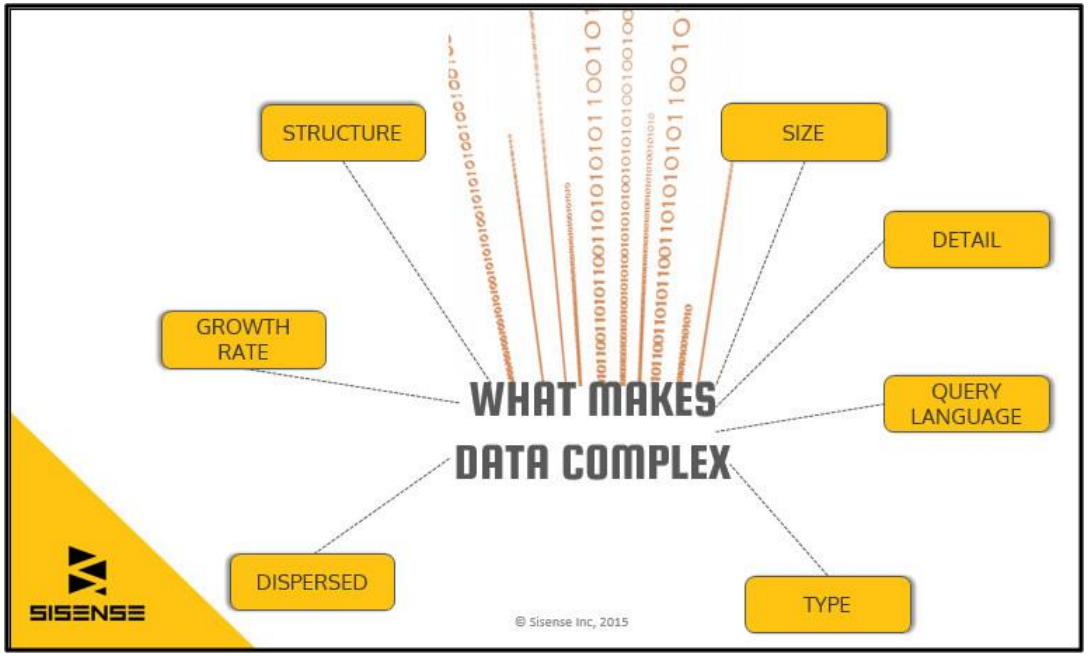
Research Area

Data analytics of complex data

- Complex data refers to information that possesses intricate characteristics, making it challenging to manage, analyze, or interpret using conventional data processing methods.
- Complexity in data can arise from various sources, including its structure, volume, diversity, and dynamics.
- Social media platforms generate vast and intricate datasets, often referred to as complex data. This complexity arises from the multifaceted nature of information shared on these platforms.

a) Sentiments Analysis

b) Recommender System



Research Area

Sentiments Analysis

Sentiment Analysis, or opinion mining, is the process of extracting emotional insights from complex data, categorizing it as positive, negative, or neutral.

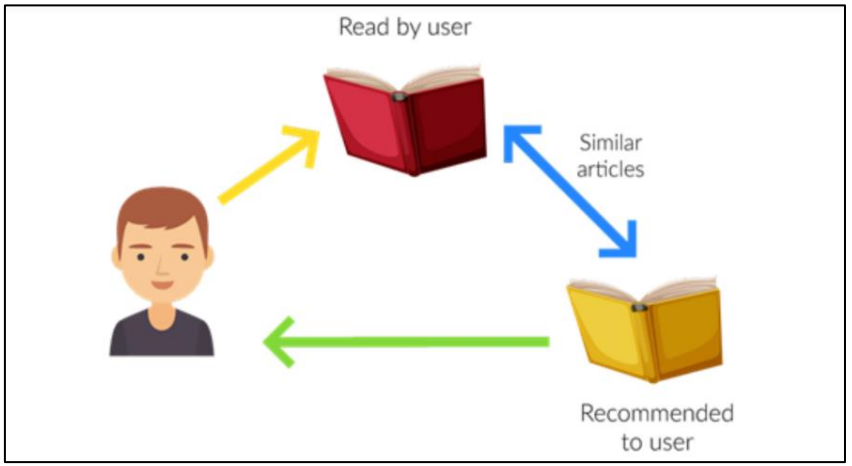
Why It Matters:

- **Customer Insights**: Uncover customer opinions, preferences, and pain points.
- **Business Decisions**: Make data-driven decisions in marketing, product development, and customer service.
- **Social Media Impact**: Monitor real-time public sentiment on social platforms.

Research Area

Recommender System

Recommender Systems, also known as recommendation systems or engines, are algorithms that provide personalized content or product suggestions to users based on their preferences and behaviors.



Applications:

- **E-Commerce**: Product recommendations in online stores.
- **Streaming Services**: Suggesting movies, music, or content.
- **Social Media**: Recommending connections, posts, or content.

Research Results

Sentiment and Emotion: Social media data is rife with sentiment and emotional expressions, providing insights into public opinions and reactions.

AI and Complex data:

- AI algorithms, particularly in machine learning and deep learning, excel in deciphering intricate patterns, recognizing relationships, and unveiling insights within diverse and multifaceted datasets.
- Whether it's cleansing and preprocessing data, understanding unstructured text through natural language processing, or analyzing complex images and videos, AI's capabilities transcend the challenges posed by complex data.
- Moreover, AI empowers applications like recommendation systems, anomaly detection, and predictive modeling in a data landscape characterized by high dimensionality, temporal dynamics, and a multitude of data types.
- By leveraging AI, we unlock the latent value within complex data, enhancing decision-making, enabling personalized experiences, and driving innovation in various domains.

Products

	Journal Papers
[P1]	Umair, Areeba, and Elio Masciari. "Sentimental and spatial analysis of COVID-19 vaccines tweets." Journal of Intelligent Information Systems (2022): 1-21 (**IF=2.504, SCOPUS** and **ISI Web of Science indexed**) **(Published)**.
[P2]	Umair, A., Masciari, E. Habib Ullah, M. Vaccine Sentimental Analysis using BERT+NBSVM and Geo-Spatial Approaches, Journal of supercomputing (2022). **IF=2.557, SCOPUS** and **ISI Web of Science indexed**) **(Published)**.
[P3]	Umair, Areeba, and Elio Masciari, Sentiment Analysis using Improved CT-BERT_CONVLayer Fusion Model for COVID-19 Vaccine Recommendation, Journal, 2023 (to be submitted)
	Conference Papers
[P4]	Areeba Umair , Elio Masciari, and Muhammad Habib Habib Ullah, Sentimental analysis applications and approaches during covid-19: a survey, Proceedings of the 25th International Database Engineering and Applications Symposium.2021.

Products

[P5]	Umair, Areeba, and Elio Masciari. "A Survey of Sentimental Analysis Methods on COVID-19 Research." SEBD (2022) **(Published) (SCOPUS indexed)**.
[P6]	Areeba Umair, and Elio Masciari, Using high performance approaches to covid-19 vaccines sentiment analysis, 2022 30th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP). IEEE, 2022.
[P7]	Areeba Umair, Elio Masciari, Sentimental Analysis of COVID-19 Vaccine Tweets Using BERT+ NBSVM, Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer Nature Switzerland, 2022.

Ph.D. Thesis Overview

"Considering COVID-19 as an example of complex data"

The social media has a vast amount of user-generated complex data.

Sentiment analysis is the field where people's feeling are extracted.

COVID-19 pandemic has affected people's lives all over the globe.

It caused the feelings of fear, anxiety, anger, depression and other issues.

Ph.D. Thesis

The review of thirty primary studies has been conducted.
The purpose of conducting this review was to explore and identify:

Benchmark data sets and well-known data sources

The volume of data used in individual study.

The types of approaches or techniques.

Application areas of COVID research

Future trends.

Ph.D. Thesis

Results of the review/survey

Data Sources	• Twitter • Online media and forums
Approaches	• Machine learning approaches, lexicon-based approaches, and hybrid approaches. • Naive Bayes and SVM
Application Areas	• Students' mental health • Reopening sentiments,
Future trends	• Explore public trust and confidence in existing policies. • More specific topics can be analyzed to help policy maker.

Ph.D. Thesis

PROBLEM Statement: The main challenges lie in efficiently processing large-scale, unstructured text data, capturing nuanced emotions, opinions, and context expressed by individuals, and effectively recommending vaccines by considering dynamic vaccine distribution, efficacy information, and individual health profiles.

- Existing sentiment analysis models may struggle to comprehend the complexity of emotions and language used during the pandemic.

- Traditional recommender systems often lack the ability to adapt to dynamic vaccine distribution scenarios, individual health conditions, and vaccine efficacy data.

Ph.D. Thesis

OBJECTIVES:

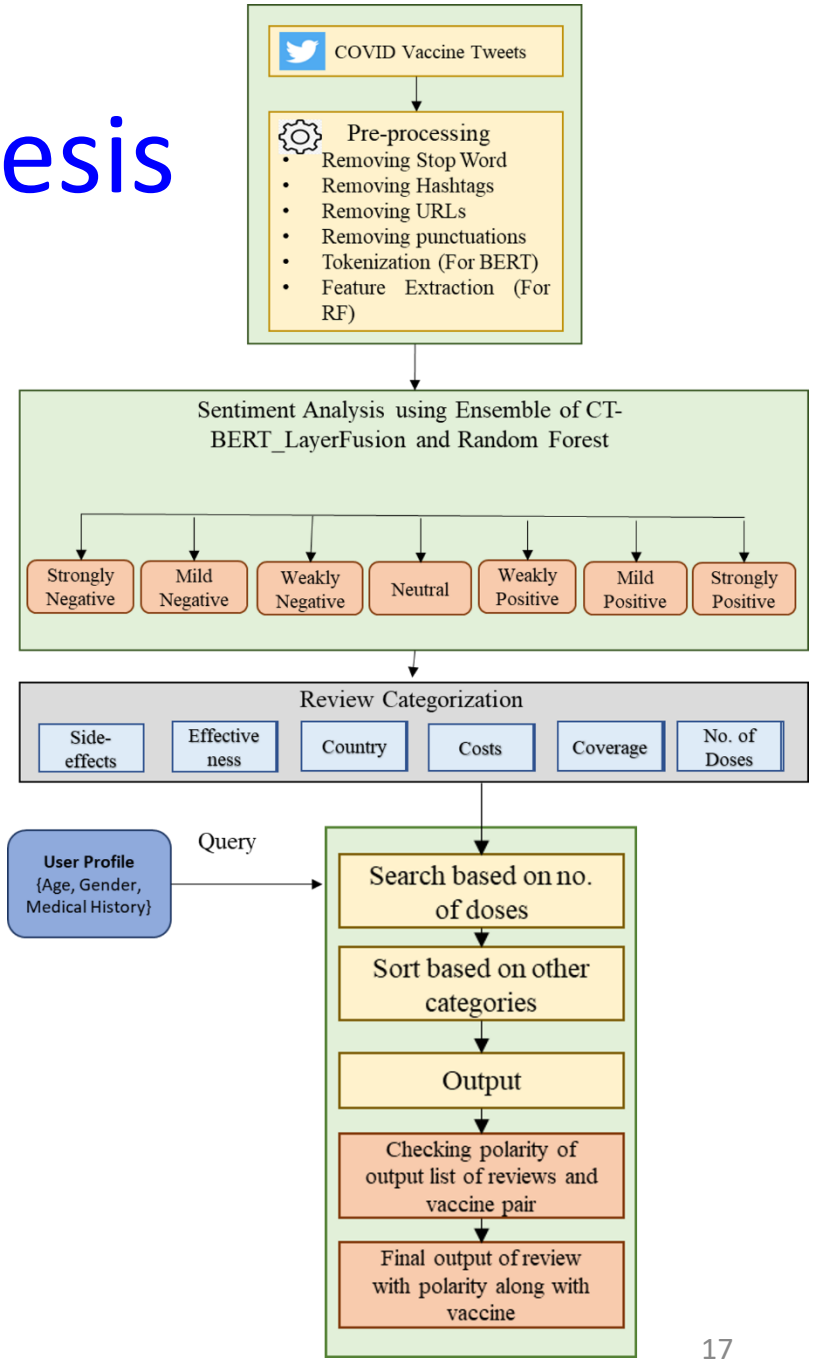
The primary objective of this research is to develop:

- ❑ Sophisticated AI tools
- ❑ Handle complex COVID-19 data
- ❑ Sentiment analysis (focusing on seven categories of tweets)
- ❑ Vaccine recommendation.

Ph.D. Thesis

METHODOLOGY:

- Vaccine Recommendation System
- Based on sentiments analysis
- Using ensemble approach
- Proposed CT-BERT-CONVLayer_Fusion



Ph.D. Thesis

Sentiments Classes:

In traditional sentiment analysis models, text data is typically categorized into three primary sentiment classes:

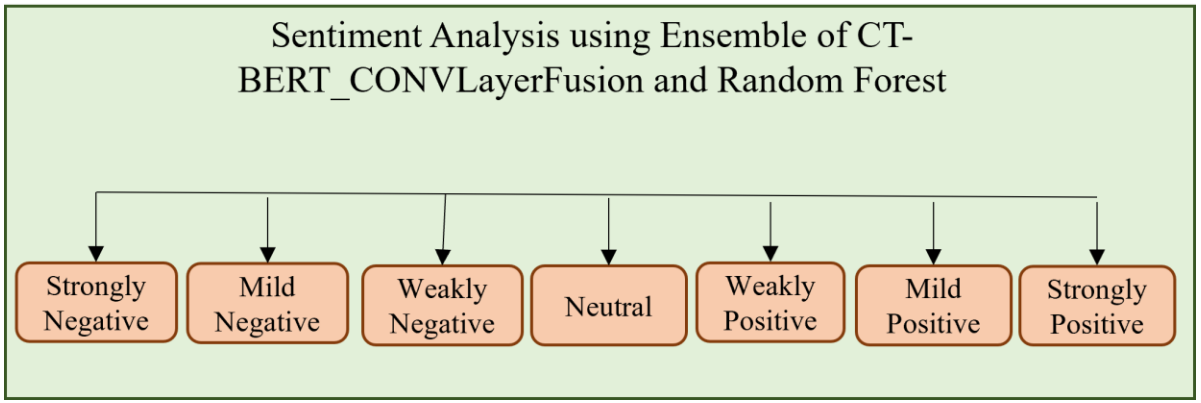
1.Positive

2.Negative

3.Neutral

Contribution:

Categorizing text into seven sentiment classes provides a more granular understanding of sentiment in language and allows for a more nuanced analysis of opinions and emotions expressed in text data.

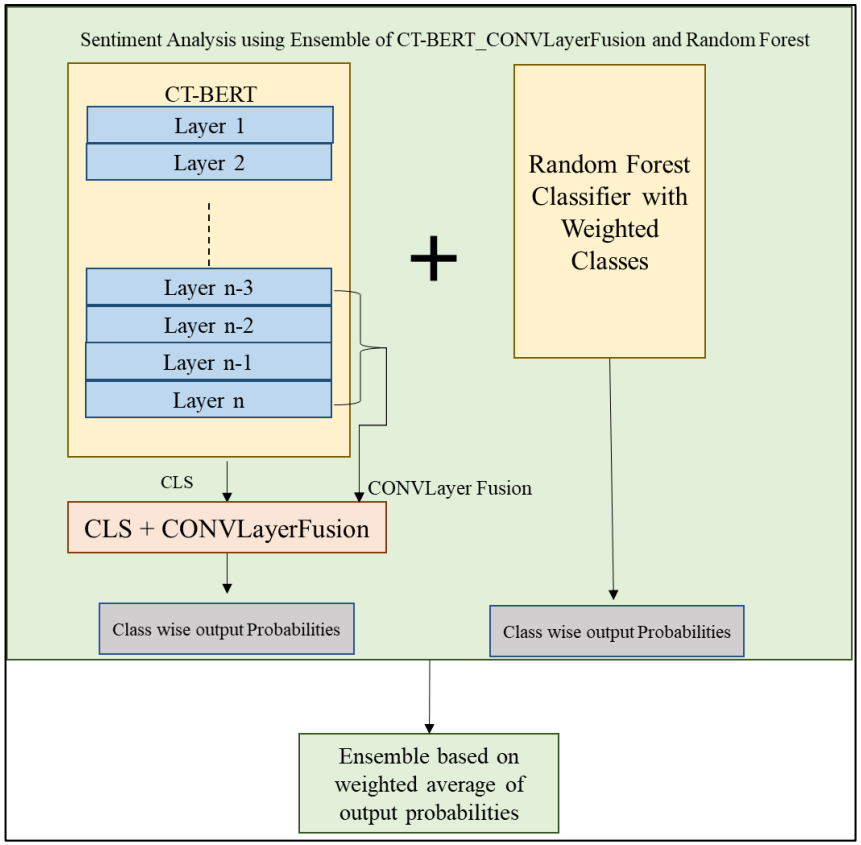


Ph.D. Thesis

Sentiment Analysis using Ensemble approach combining CT-BERT_CONVLayerFusion with a Random Forest classifier.

METHODOLOGY:

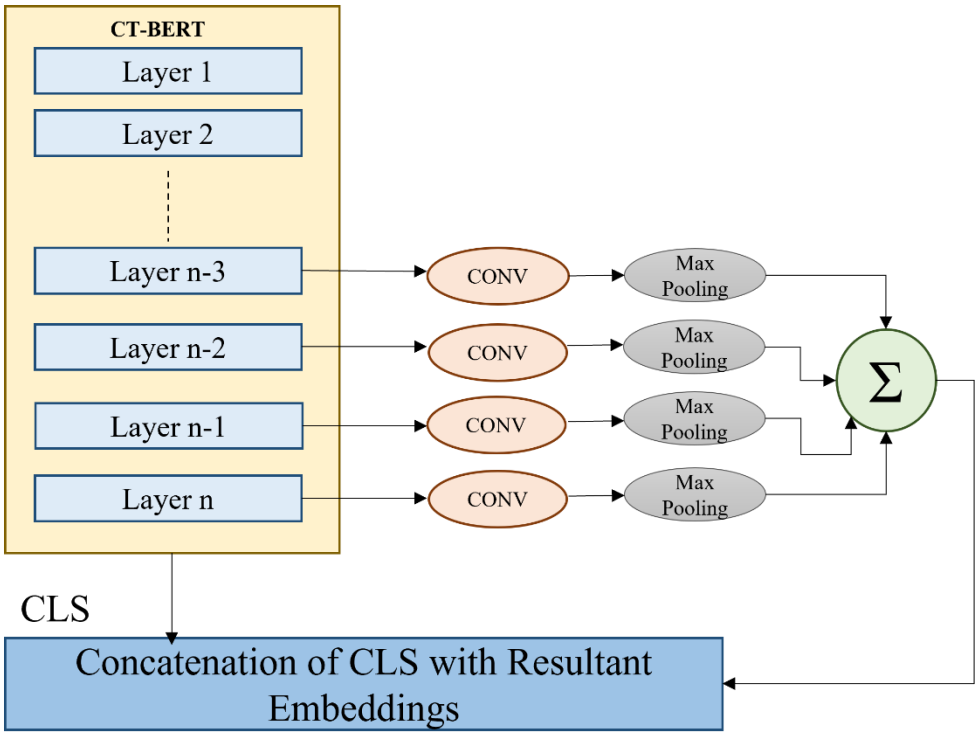
- The CT-BERT model, is a transformer-based model developed for COVID-19 tweets.
- To enhance the CT-BERT model, the last four layers were improved by incorporating convolutional layers.



Ph.D. Thesis

Main Contributions:

- Following the convolutional layers, the MAX Pooling function was applied individually on each layer.
- After applying maximum pooling, the resulting embeddings from each layer were stacked.
- These stacked embeddings were then summed.



Ph.D. Thesis

The Proposed Algorithm:

Algorithm 1 CT-BERT-LayerFusion

1: 1: Strongly Negative Tweets
2: 2: Mild Negative Tweets
3: 3: Weakly Negative Tweets
4: 4: Neutral Tweets
5: 5: Weakly Positive Tweets
6: 6: Mild Positive Tweets
7: 7: Strongly Positive Tweets
8: Conv(): Convolution layer
9: CT-BERT_Max(): Maximum value embeddings of each layer of CT-BERT
10: CT-BERT_MaxLayersSum(): Sum of Pooled embeddings
11: D: Dataset
12: Input D
13: Steps:
14: n= CT-BERT_Layers
15: **for** i=n-3 to n **do**
16: Conv()
17: Conv_Max()
18: **end for**
19: **for** i = 1 to size of (D) **do**
20: Final_Embeddings ← Concatenation(CLS(D_k), CT-BERT_MaxLayersSum(D_k)
21: Tweet_Classification ← Classifier(Final_Embeddings)
22: **end for**
23: Output: Tweet_Classification (Class wise probabilities)

Ph.D. Thesis

- In the proposed approach, for each of the last four transformer layers, the output tensor was extracted.
- After obtaining the output tensor, a convolutional layer was added.
- The configuration of the convolutional layer involved setting parameters such as kernel size, stride, padding, and the number of filters.
- To ensure compatibility between the transformer layer and the convolutional layer, the input dimensions of the convolutional layer were adjusted to match the output dimensions of the corresponding transformer layer.
- This alignment of dimensions was crucial to maintain consistency in the flow of information between the layers.

Ph.D. Thesis

Tweet Categorization

In the categorization process, tweets or reviews are grouped into pre-defined categories based on the most frequent words found in the dataset.

To categorize the reviews, two different processes are employed:

- Fuzzy string matching
- Angular similarity.

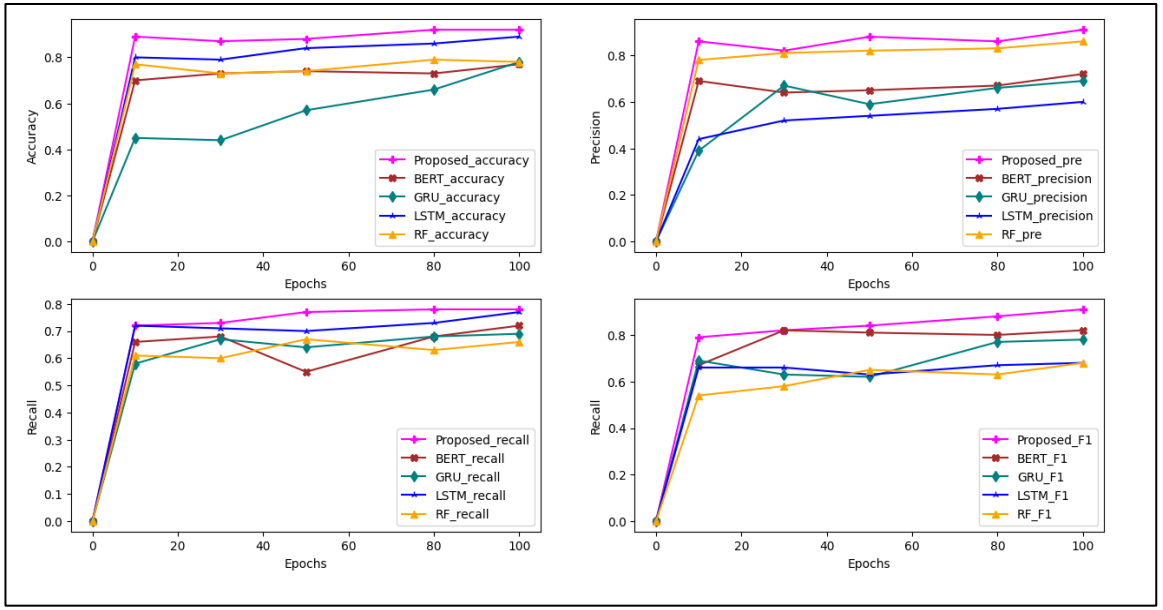
These processes compare the similarity between each review and the index terms in different categories.

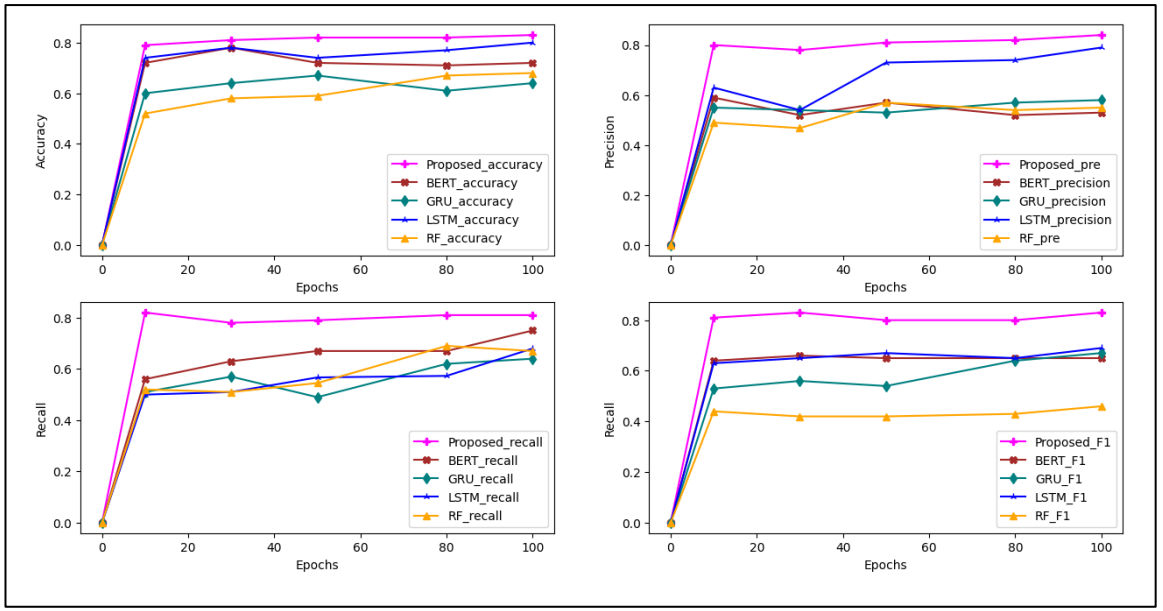
Ph.D. Thesis Results

RESULTS:

The proposed method outperformed other models and achieved 88 % accuracy, 82 % precision, 78 % recall and 82 % F-measure for classification of strongly positive sentiments while 80 % accuracy, 76 % precision, 81 % recall and 83 % F-measure for classification of strongly negative sentiments respectively.

Results of 7 sentiment classes are given in Thesis.





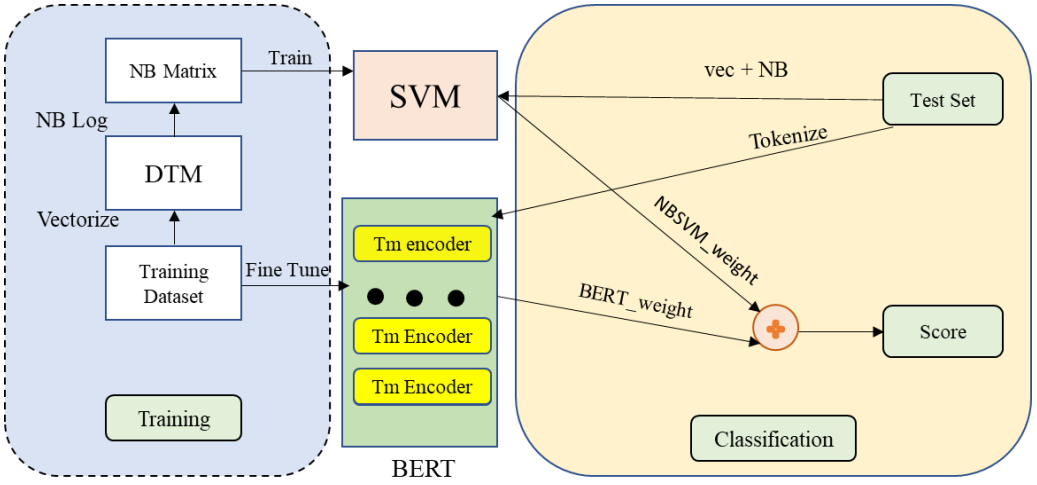
Thank you for the attention

Ph.D. Thesis Part 1

Sentiment analysis of COVID-19 Vaccines tweets using BERT+NBSVM model

METHODOLOGY:

- BERT+NB-SVM model is estimated on DTM (document term frequency) features.
- The training dataset was used in fine tuning of BERT model.
- The DTM is used to compute the NB log-count ratios.
- The NB log-count ratios are used for SVM model training
- While prediction, final score is calculated as the sum of the fitted NB-SVM model and best fine-tuned BERT model.
- The best fine-tuned model indicates the model with best performances with different epochs and batch sizes.

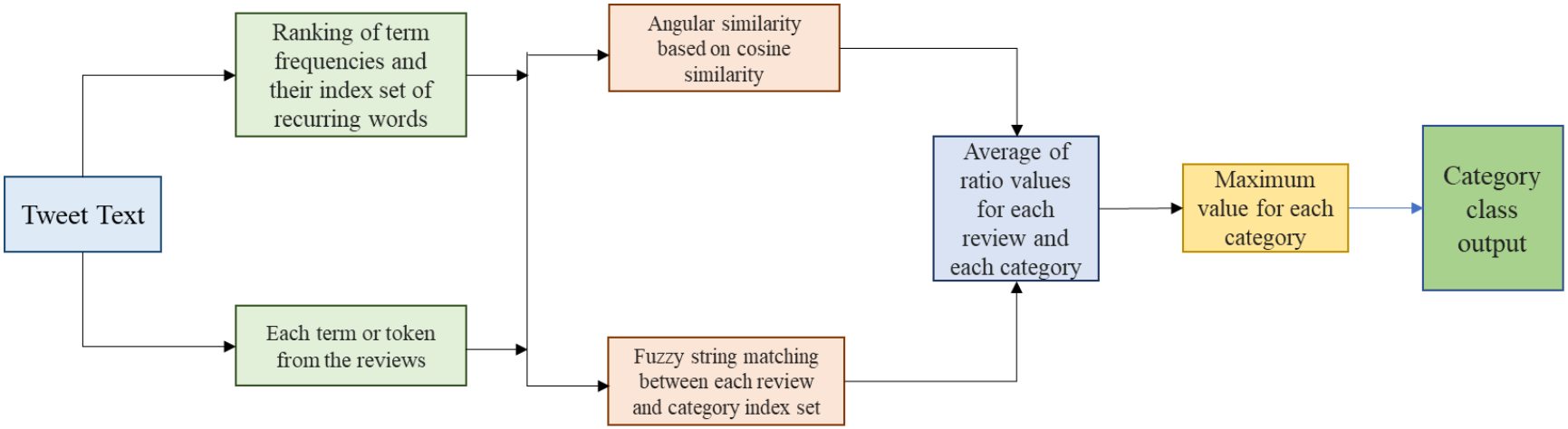


RESULTS:

The proposed BERT+NBSVM outperformed other models and achieved 73 % accuracy, 71 % precision, 88 % recall and 73 % F-measure for classification of positive sentiments
while 73 % accuracy, 71 % precision, 74 % recall and 73 % F-measure for classification of negative sentiments respectively.

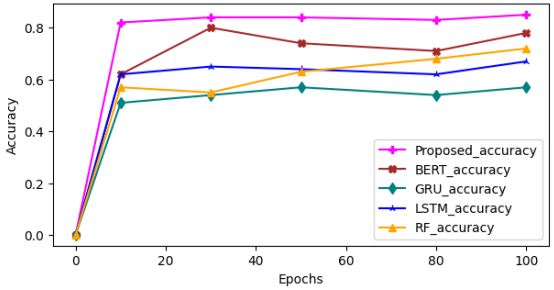
PhD thesis-Appendix

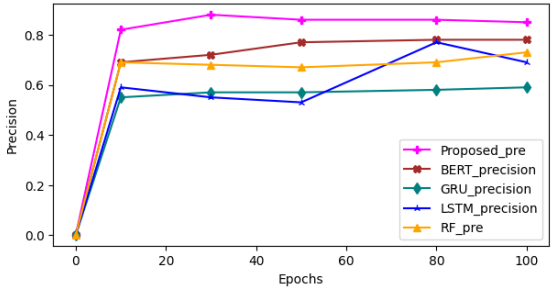
Tweet Categorization

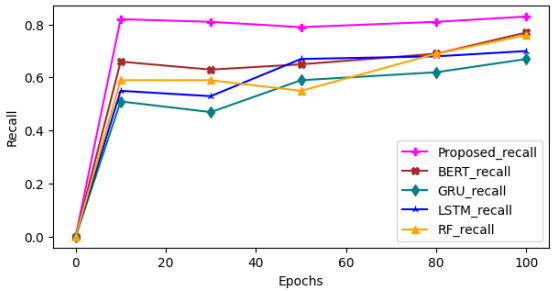


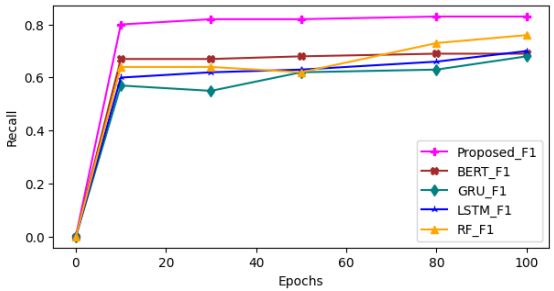
PhD thesis-Appendix

Mild Negative

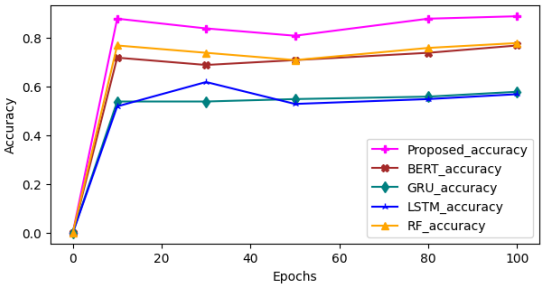


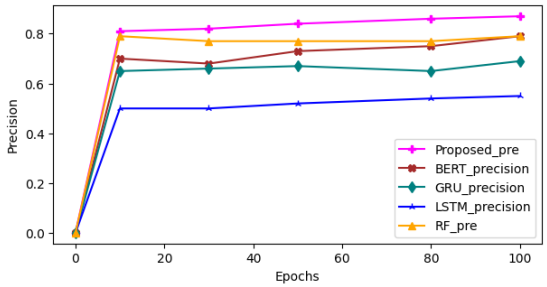


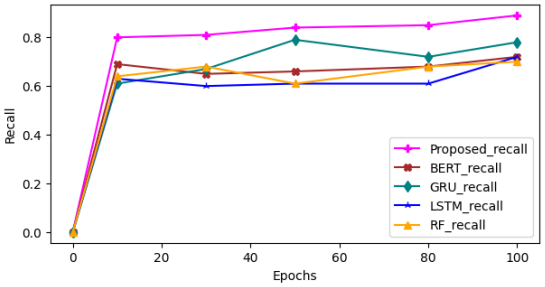


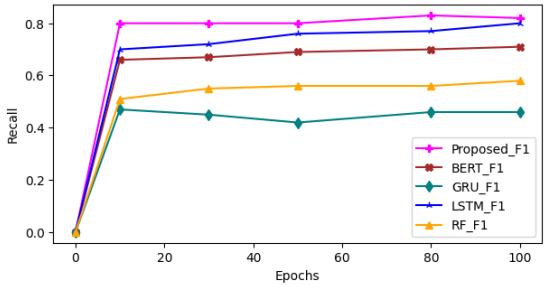


Weakly Negative



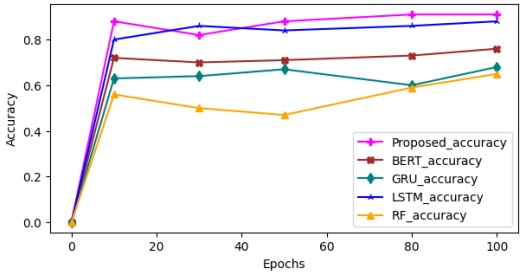


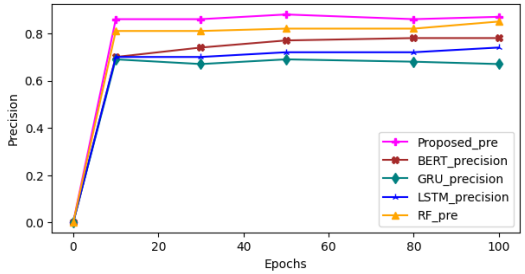


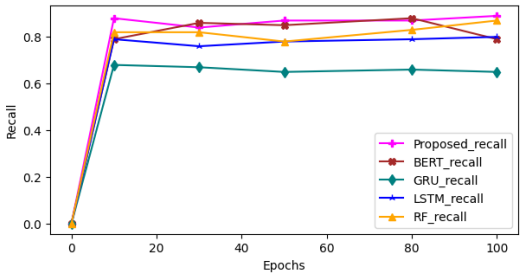


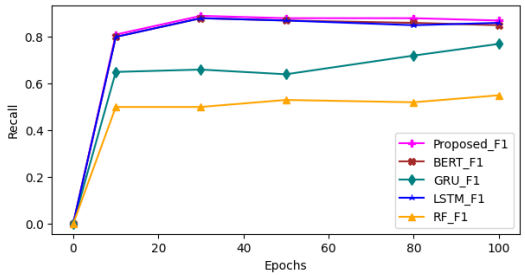
PhD thesis-Appendix

Mild Positive

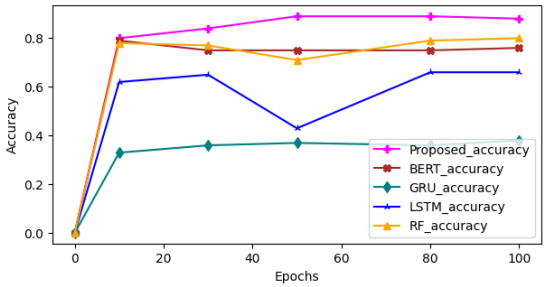


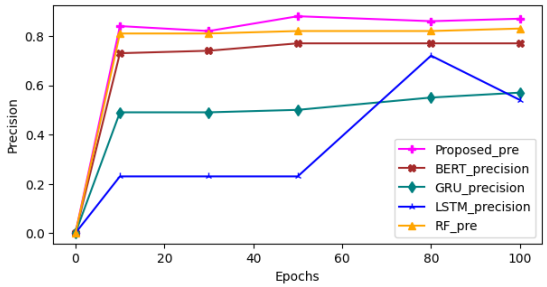


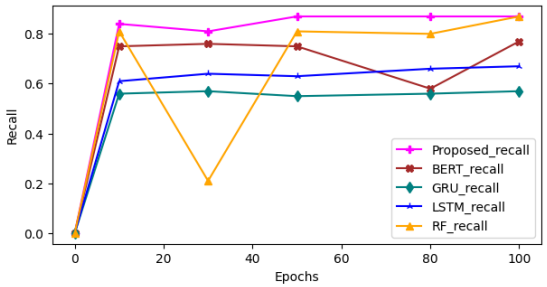


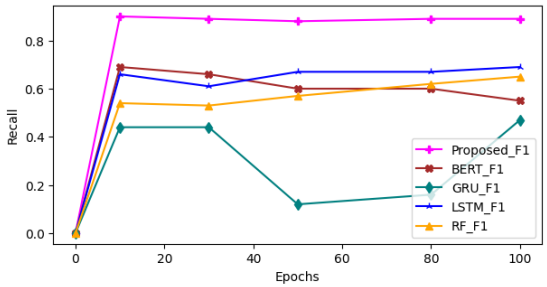


Weakly Positive

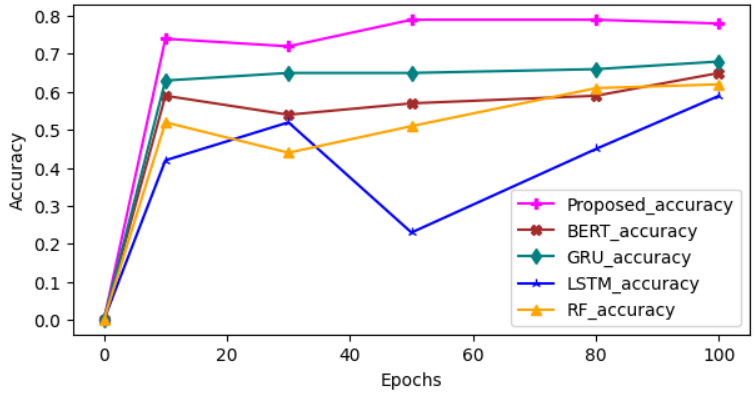


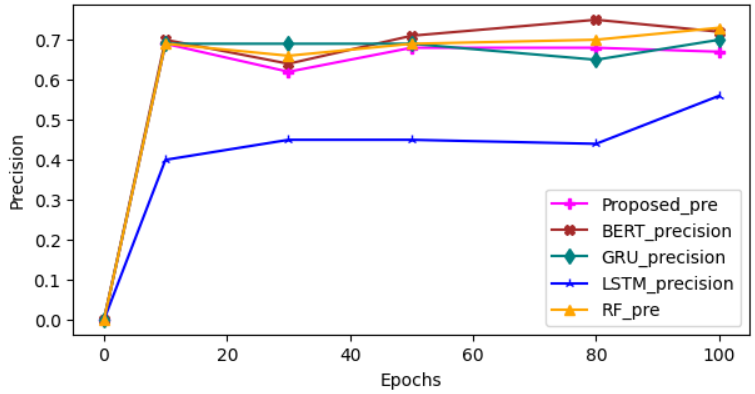


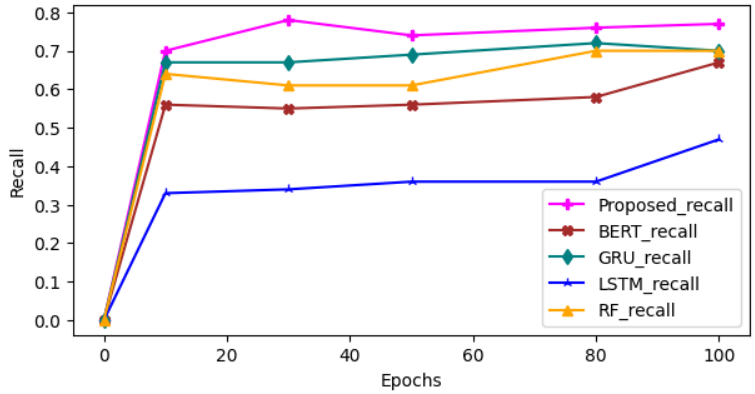


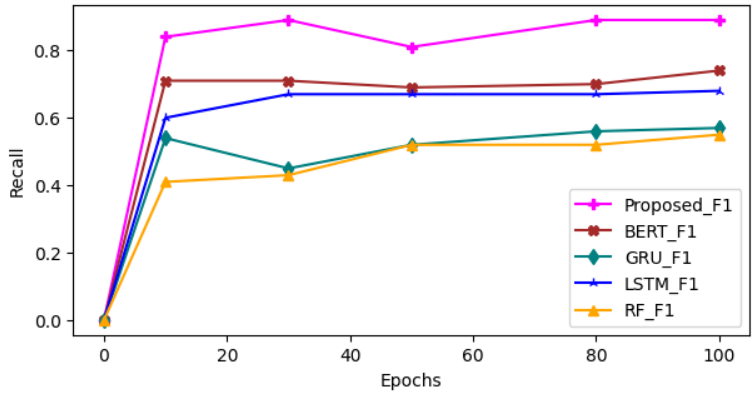


Neutral


Accuracy
Epochs
Proposed_accuracy
BERT_accuracy
GRU_accuracy
LSTM_accuracy
RF_accuracy


Precision
Epochs
Proposed_pre
BERT_precision
GRU_precision
LSTM_precision
RF_pre


Recall
Epochs
Proposed_recall
BERT_recall
GRU_recall
LSTM_recall
RF_recall


Recall
Epochs
Proposed_F1
BERT_F1
GRU_F1
LSTM_F1
RF_F1

Appendix

- Sigma Architecture was used to Fine Tunning of BERT using High Performance architecture