Università degli Studi di Napoli Federico II
Ph.D. Program in
Information Technology and Electrical Engineering
XXXVI Cycle

Thesis for the Degree of Doctor of Philosophy

Devising Artificial Intelligence Tools For Complex Data

by

Areeba Umair

Advisor: Prof. Elio Masciari

Scuola Politecnica e delle Scienze di Base
Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione

This thesis is dedicated to my loving husband, my parents and my siblings.

Devising Artificial Intelligence Tools For Complex Data

Ph.D. Thesis presented
for the fulfillment of the Degree of Doctor of Philosophy
in Information Technology and Electrical Engineering
by

Areeba Umair

October 2023

Approved as to style and content by

____

Prof. Elio Masciari, Advisor

Università degli Studi di Napoli Federico II
Ph.D. Program in Information Technology and Electrical Engineering
XXXVI cycle - Chairman: Prof. Stefano Russo

http://itee.dieti.unina.it

Candidate's declaration

I hereby declare that this thesis submitted to obtain the academic degree of Philosophiæ Doctor (Ph.D.) in Information Technology and Electrical Engineering is my own unaided work, that I have not used other than the sources indicated, and that all direct and indirect sources are acknowledged as references.
Parts of this dissertation have been published in international journals and/or conference articles (see list of the author's publications at the end of the thesis).

Napoli, December 14, 2023

Areeba

Areeba Umair

Abstract

Complex data, in the context of data science, refers to information with complex structures, such as high-dimensionality, mixed data types, temporal or spatial dependencies, graph formats, unstructured content, missing values, nonlinear relationships, hierarchical organization, or dynamic changes over time. Social media data is considered complex due to its diverse formats, high volume, unstructured nature, temporal dynamics, network structure, noise, missing data, sentiment, and contextual challenges. Social media dataset encompasses a range of information, including public reactions, opinions, news, and discussions. Analyzing this data provides insights into public sentiment, and thus also help to adopt strategies and campaigns according to people's need.

Various techniques have been employed for COVID-19 sentiment analysis including Lexicon-based methods, supervised machine learning models, Deep learning approaches, transformer-based models, Ensemble methods, and aspect-based analysis. In the first phase of this thesis, we used freely available X app (former Twitter) complex data and proposed BERT+NBSVM for classifying negative and positive tweets regarding COVID-19 vaccines, after applying necessary pre-processing steps. In second phase of thesis, we proposed sentiment analysis based recommender system for COVID-19 vaccines. For this purpose, we proposed an ensemble of random forest with CT-BERT_CONVLayerFusion model, for classifying the tweets into seven different categories of sentiments. We also utilized some of the Geo-Spatial approaches to geographically analyse the peoples sentiments. The proposed techniques have shown encouraging results from both a qualitative and quantitative point of view. All the results are published in reputed Journals and International Conferences.

Keywords: Complex data, Artificial Intelligence, Recommender System, Sentiment Analysis, BERT.

Sintesi in lingua italiana

Per complessità dei dati, nell'ambito della Data Science, ci si riferisce a informazioni con strutture complesse, come ad esempio dati ad alta dimensionalità, dati eterogenei, dipendenze temporali o spaziali, dati a grafo, contenuti non strutturati, valori mancanti, relazioni non lineari, organizzazione gerarchica o cambiamenti dinamici nel tempo. I dati provenienti dai social media, sono considerati complessi a causa dei loro formati diversificati, del loro elevato volume, della loro natura non strutturata e della dinamicità temporale degli stessi. Questo insieme di dati comprende una serie di informazioni, tra cui reazioni pubbliche, opinioni, notizie e discussioni su vari aspetti della pandemia che sono state ampiamente utilizzate per la redazione della presente tesi. L'analisi di questi dati pertanto, fornisce intuizioni sulle opinioni degli utenti, le tendenze nei fenomeno osservati, la possibile influenza della disinformazione e aiuta anche a fornire suggerimenti appropriati in ambiti specifici e cruciali come quello sanitario.

In questa tesi, abbiamo utilizzato i dati complessi dell'app X (ex Twitter) e proposto l'utilizzo di BERT+NBSVM per classificare i tweet negativi e positivi riguardanti i vaccini COVID-19, dopo aver applicato le necessarie fasi di pre-elaborazione. A tale scopo, è stato proposto un approcio basato su random forrest con il modello CT-BERT_CONVLayerFusion, per classificare i tweet in sette diverse categorie di sentimenti. Abbiamo anche utilizzato alcuni approcci geospaziali per analizzare geograficamente il sentiment delle persone.

Le tecniche proposte hanno mostrato risultati incoraggianti sia dal punto di vista qualitativo che quantitativo e ottenuto un buon riscontro anche nelle pubblicazioni che ne sono scaturite.

Parole chiave: Dati complessi, Intelligenza Artificiale, Recommender System, Sentiment Analysis, BERT.

Contents

Abstract i

Sintesi in lingua italiana ii

Acknowledgements vii

List of Acronyms x

List of Figures xiii

List of Tables xv

1 Introduction **1**

1.1 Handling complex data 2

1.2 Relationship between Big Data and Complex Data 3

1.2.1 Big Data: The Volume, Velocity, and Variety Paradigm 4

1.2.2 Complex Data: Navigating Intricacies within Datasets 4

1.2.3 The Symbiotic Relationship 4

1.3 COVID-19 Twitter Data as a case study of complex data . . 5

1.4 Sentiments Analysis of complex data 6

1.5 Recommender System for complex data 7

1.6 Geographic Information System and complex data 8

1.7 Thesis Structure 9

2 Background and Related Work **11**

2.1 Methodology 11

2.1.1 Data Sources during COVID-19 research: 12
2.1.2 Approaches for COVID-19 Sentimental Classification 12
2.1.3 Sentimental Analysis Applications during COVID-19 14
2.2 Comparison of Studies 16
2.3 State-of-the-art-algorithms 22

3 Improving AutoEncoder for Sentiments Classification 25
3.1 Research Questions 25
3.2 Methodology 26
3.2.1 Proposed Framework 26
3.2.2 Collection of data 26
3.2.3 Getting Sentiment Polarity Values 28
3.2.4 Word Clouds 29
3.2.5 Generic BERT for Sentimental Classification 30
3.2.6 Combining BERT and Naive Bayes-SVM for Sentimental Classification 33
3.3 Achieving High Performances: The Sigma Architecture . . . 35
3.4 Results and discussion 37
3.4.1 Evaluation Matrices 37
3.4.2 Sentiment Polarity Computing 38
3.4.3 RQ1: What are the key concerns and sentiments expressed by individuals in the analyzed X app data regarding COVID-19 vaccination? 38
3.4.4 RQ2: How effective is the proposed sentiment analysis model in classifying and visualizing sentiments from X app (former Twitter) data? 40

4 Recommendation System for COVID-19 Vaccines 43
4.1 Research Questions 43
4.2 Proposed Recommendation System 43
4.2.1 Sentiment Analysis 44

4.2.2 CT-BERT (COVID Twitter BERT) Model 47
4.2.3 Random Forest 51
4.2.4 Tweets Categorization 52
4.3 Results and Discussion 55
4.3.1 Comparison With the State-of-the-Art and Evaluation Matrices 56
4.3.2 RQ1: How does the proposed ensemble approach, combining CT-BERT-CONVLayer_Fusion and Random Forest, contribute to accurate sentiment classification of COVID-19-related tweets? 56
4.3.3 RQ2: To what extent does the sentiment analysis contribute to the categorization of COVID-19 vaccine-related tweets into specific predefined categories, and how does this enhance the recommendation system's comprehensiveness? 63
4.4 Main Contributions 64

5 Geo-Spatial Analysis of COVID Vaccine Tweets 67
5.1 Methods 67
5.1.1 Vaccine hesitancy due to access issues. 68
5.1.2 Geo-Coding and Visualization of Data: 68
5.1.3 Geographical correlation: 69
5.1.4 Hotspot Analysis 71
5.1.5 Analysis Using Kernel Density Estimation: 72
5.1.6 Buffering 73

6 Conclusions 75
6.1 Limitations and Future Works 77

Author's Publications 79

Bibliography **81**

Acknowledgements

The research presented in this dissertation has been supported by the Italian Ministry of University and Research (MUR) under Ph.D. program 2020-2023

List of Acronyms

The following acronyms are used throughout the thesis.

ML	Machine Learning
NLP	Natural Language Processing
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
NB	Naive Bayes
SVM	Support Vector Machine
LSTM	Long-Short Term Memory
DT	Decision Tree
LDA	Latent Dirichlet Allocation
LR	Linear Regression
LoR	Logistic Regression
KNN	K-Nearest Neighbour
URL	Uniform Resource Locators

MLM	Maked Language modelling
CLS	Classification
SEP	Seperation
BERT	Bidirectional Encoder Respresentation of Trasnforemrs
NBSVM	Hybrid of Naive Bayes and Support Vector Machine
DTM	Document Term Matrix
TN	True Negative
TP	True Positive
FN	False Negative
FP	False Positive
RF	Random Forest
GIS	Geographical Information System

List of Figures

3.1 The overall proposed methodology of our research 26

3.2 Threshold of polarity for sentiment classes. 30

3.3 How TextBlob() works? 30

3.4 Architecture of BERT model 32

3.5 System architecture of BERT+ NBSVM 34

3.6 Sigma Architecture 36

3.7 Figure contains Positive word cloud of tweets, Negative word cloud of tweets and the neutral word cloud of tweets 39

3.8 Comparison of BERT+NBSVM model with state-of-the-art for positive sentiment classification. 40

3.9 Comparison of BERT+NBSVM model with state-of-the-art for negative sentiment classification. 41

4.1 Architecture of proposed recommendation system for COVID-19 vaccines 46

4.2 Architecture for the proposed ensemble of CT-BERT-CONVLayerFusion with a Random Forest classifier 47

4.3 Proposed CONV Layer Fusion 49

4.4 Wordcloud for determining the frequent words in vaccine review data 54

4.5 A review with the similarity values with each of the category 55

4.6 Architecture for the review categorization process 55
4.7 Graph of Training and Validation accuracy of our proposed the CT-BERT_CONVLayerFusion model 57
4.8 Sub-graphs of Accuracy, Precision, Recall and F1 showing the comparison of our proposed Sentiment Analysis model with state-of-the-art for strongly negative tweet classification. 58
4.9 Sub-graphs of Accuracy, Precision, Recall and F1 showing the comparison of our proposed Sentiment Analysis model with state-of-the-art for mild negative tweet classification. . 59
4.10 Sub-graphs of Accuracy, Precision, Recall and F1 showing the comparison of our proposed Sentiment Analysis model with state-of-the-art for weakly negative tweet classification. 59
4.11 Sub-graphs of Accuracy, Precision, Recall and F1 showing the comparison of our proposed Sentiment Analysis model with state-of-the-art for neutral tweet classification. 60
4.12 Sub-graphs of Accuracy, Precision, Recall and F1 showing the comparison of our proposed Sentiment Analysis model with state-of-the-art for weakly positive tweet classification. 60
4.13 Sub-graphs of Accuracy, Precision, Recall and F1 showing the comparison of our proposed Sentiment Analysis model with state-of-the-art for mild positive tweet classification. . 61
4.14 Sub-graphs of Accuracy, Precision, Recall and F1 showing the comparison of our proposed Sentiment Analysis model with state-of-the-art for strongly positive tweet classification. 61
4.15 Pie Chart showing the distribution of tweets based on the category type 64

5.1 Vaccine data Visualization 69
5.2 Spatial correlation among the data features 71
5.3 Hotspot analysis of COVID-Vaccine tweets 72

5.4 Kernel Density Estimation of COVID-Vaccine tweets 73
5.5 Results of Buffering on Vaccination dataset 74

List of Tables

2.1 Sentimental Analysis Approaches and Applications During COVID-19 Pandemic 17

3.1 Comparison of tweets before and after pre-processing 29

3.2 Parameters used for training the model 35

3.3 Polarity values and sentiment categories with respect to sample tweets 38

4.1 Top 20 features and their importance obtained from random forest classifier. 53

4.2 Hyperparameters for Sentiment Classification 56

5.1 Confidence level associated with P-values and Z-scores . . . 70

Chapter 1

Introduction

Complex data refers to information that exhibits structures, relationships, and features that cannot be dealt using classical approaches, thus making it challenging to analyze, process, or interpret using traditional methods or tools. Complex data can arise in various domains, including scientific research, finance, healthcare, social networks, and artificial intelligence applications.

Characteristics of complex data may include:

1. High Dimensionality: Data with a large number of features or variables can lead to the "curse of dimensionality," where traditional algorithms struggle to handle the exponential growth of computational resources required.

2. Heterogeneity: Complex data often consists of different data types or formats, such as numerical, categorical, text, and image data, which require specialized techniques to integrate and analyze effectively.

3. Sparsity: In some datasets, a significant portion of the information may be missing, leading to sparse data matrices that require specialized algorithms to deal with missing values.

4. Non-Linearity: The relationships between variables may not follow linear patterns, requiring more sophisticated models that can capture non-linear interactions.

5. Temporal or Sequential Dependencies: Data may exhibit dependencies over time or in sequences, requiring specialized time series analysis or sequential modeling techniques.

6. Noisy Data: Complex data can be noisy, containing errors or outliers that can affect analysis and modeling.

7. Imbalance: In certain datasets, classes or categories may be imbalanced, with some categories having a disproportionately low number of samples compared to others.

8. Homophily: This concept is particularly relevant when data is structured as graphs, reflecting relationships characterized by similarity or affinity. Homophily suggests that entities within a network tend to associate with others who share similar attributes or characteristics. In the context of complex data represented as graphs, this phenomenon shapes the dynamics of connections, fostering clusters of similarity.

1.1 Handling complex data

Handling complex data often requires the use of advanced algorithms and techniques, such as artificial intelligence, deep learning, data dimensionality reduction methods, ensemble learning, and specialized data preprocessing methods. Additionally, domain expertise and collaboration between data scientists, domain experts, and stakeholders are vital to effectively analyze and extract insights from complex datasets. Complex data plays a significant role in the field of artificial intelligence (AI). As AI aims to develop systems that can mimic human intelligence and decision-making, it often encounters complex datasets that require specialized techniques to be effectively processed and analyzed. Here's how complex data and AI are interconnected:

1. Model Complexity: Complex datasets often demand more sophisticated AI models to accurately capture intricate patterns and relationships. For example, deep learning models, such as deep neural

networks, are capable of handling complex data, such as images, audio, and natural language, by learning hierarchical representations of the data.

2. Natural Language Processing (NLP): NLP deals with the complexities of human language, such as ambiguity, context, and syntax. AI-powered language models like GPT-3 have demonstrated the ability to process and generate human-like text, enabling applications like language translation, sentiment analysis, and chatbots.

3. Unstructured Data: Complex data often includes unstructured data such as social media data. AI techniques, such as word embeddings, feature extraction, are employed to extract valuable insights from this unstructured data.

4. AI and Complex data: In the realm of complex data, the need for explainability in AI results is paramount. As datasets become increasingly intricate, characterized by interconnected relationships and diverse patterns, understanding the decisions made by AI models becomes more challenging yet crucial. Explainability serves as a critical tool to unravel the intricacies of these complex datasets, providing transparency into the decision-making processes of AI algorithms.

Overall, the ability of AI to effectively handle complex data is crucial for its widespread adoption and successful application across various domains, bringing about advancements and improvements in many aspects of modern life.

1.2 Relationship between Big Data and Complex Data

In the contemporary landscape of information technology, the realms of "big data" and "complex data" stand as pivotal cornerstones, shaping the way we comprehend, analyze, and extract insights from vast datasets. The advent of big data, characterized by the immense volume, velocity, and variety of information, has revolutionized our capacity to process and derive value from data sources. Simultaneously, the intricacies embedded within the data itself, commonly referred to as complex data, introduce an additional layer of challenge and opportunity.

1.2.1 Big Data: The Volume, Velocity, and Variety Paradigm

Big data, as a paradigm, encapsulates datasets that exceed the capacities of traditional data processing methods. It is distinguished by the three Vs:

1. Volume: The sheer magnitude of data generated continuously from diverse sources.
2. Velocity: The speed at which data is produced, processed, and analyzed in real-time.
3. Variety: The diversity of data types, including structured, semi-structured, and unstructured data.

The amalgamation of these three Vs necessitates advanced computational and analytical frameworks, challenging traditional data processing approaches.

1.2.2 Complex Data: Navigating Intricacies within Datasets

While big data introduces challenges in managing sheer scale, complex data delves into the intricacies woven within the information fabric. Complex data comprises data with intricate structures, interdependencies, and diverse formats. This complexity is often encountered in real-world scenarios, where data exhibits multifaceted relationships, hierarchies, and nonlinear patterns.

1.2.3 The Symbiotic Relationship

The relationship between big data and complex data is symbiotic. Big data encompasses datasets of unprecedented sizes, often inheriting complexity through its diverse variety. Conversely, complex data, irrespective of its volume, benefits from big data technologies to efficiently process, analyze, and uncover hidden patterns within its intricate structure.

1.3 COVID-19 Twitter Data as a case study of complex data

The study of COVID-19 Twitter (now X) data can serve as a fascinating case study of complex data, offering insights into various aspects such as public sentiment, information dissemination, and societal reactions during a global health crisis. Coronavirus flu spread since December 2019 from the Chinese city of Wuhan. The quick diffusion of this dangerous flu caused an infectious disease sadly known as COVID-19 [2], [28]. In March 2020, it was declared as pandemic by WHO (World Health Organization) as it infected people all over the globe.

COVID-19 Twitter data are complex due to several reasons. Firstly, they encompasses a diverse range of formats, including text, images, and videos, requiring specialized analysis approaches to extract meaningful insights from each type of content [16, 78]. Moreover, the sheer volume of data is substantial, with millions of tweets posted daily, necessitating scalable data management and analysis solutions. The unstructured nature of tweets, often composed in informal language, poses challenges in processing and interpreting the text. Additionally, the data can contain typos, abbreviations, and non-conventional terms, introducing noise and making analysis more intricate. Temporal dynamics add another layer of complexity, as tweets reflect ongoing events and can exhibit rapid shifts in sentiment and opinions as the pandemic situation evolves.

The network structure of tweet data, with users connecting through mentions, retweets, and replies, demands the analysis of user relationships and information diffusion across the network. Furthermore, the data can be influenced by false positives, misinformation, and conspiracy theories, adding complexity to discerning the truth.

Lastly, sentiment analysis is complex due to the range of emotions and opinions expressed in tweets, spanning from concerns and fear to hope and gratitude [69]. Addressing these challenges requires applying advanced data analysis methodologies, including natural language processing, emotion detection, and social network analysis, to attain an accurate and meaningful understanding of the intricacies within COVID-19 Twitter data.

1.4 Sentiments Analysis of complex data

Sentiment analysis of complex data is essential for gaining valuable insights into public opinion, customer satisfaction, and emerging trends across various domains [92]. Whether in business for brand monitoring and customer experience enhancement, in politics for understanding public sentiment, or in healthcare for monitoring public health concerns, this analytical tool enables informed decision-making and proactive strategies [20, 65]. Its versatility extends to financial markets, academia, and crisis management, making it an invaluable asset for organizations and researchers seeking to understand and respond to the nuanced sentiments expressed in today's multifaceted data landscape.

The two ways to solve sentimental classification tasks are traditional machine learning Machine Learning (ML) methods and deep learning methods. The traditional methods usually use classifiers i.e. Support Vector Machine Support Vector Machine (SVM) and Naive Bayes Naive Bayes (NB) for this purpose while in deep learning methods Recurrent Neural Network Recurrent Neural Network (RNN) and Convolutional Neural Network Convolutional Neural Network (CNN) have widely used in natural language processing Natural Language Processing (NLP) tasks.

Sentiment analysis during the COVID-19 pandemic has been a critical tool in understanding the public's emotional response to this global crisis. By analyzing vast amounts of social media posts, news articles, and online discussions, sentiment analysis has provided valuable insights into how people feel about various aspects of the pandemic, including government measures, vaccination efforts, and public health guidelines. It has helped monitor misinformation and fake news, enabling authorities to address false claims promptly. Additionally, sentiment analysis has shed light on mental health concerns, tracking fear and anxiety levels among the public, and has provided valuable feedback on the effectiveness of public health communication. By harnessing sentiment analysis, decision-makers can better tailor their responses and support measures to address the emotional needs of the population during these challenging times.

1.5 Recommender System for complex data

Recommender systems play a pivotal role in distilling valuable insights from complex data by offering personalized suggestions based on diverse user preferences and contextual information [88], [36]. Navigating multimodal data integration, dynamic shifts in user behavior [31], and ethical considerations such as privacy and fairness, these systems strive to provide accurate recommendations despite sparse or noisy datasets [40], [7]. Tackling challenges like the cold start problem and incorporating evolving information [52], recommender systems serve as indispensable tools for enhancing user experience across various domains, from e-commerce to content streaming platforms [19].

Vaccines are considered as one of the major health interventions, that are saving millions of lives every year and are cost-effective and reliable. After the declaration of COVID-19 as pandemic by WHO, many pharmaceutical companies worked hard and at an unprecedented velocity for the development of vaccines [3], [34]. The development of COVID-19 vaccines resulted in an uncompared rapid release on the market of the vaccines [49],[3], [34]. In preclinical development, there are 184 vaccines while 104 vaccines are in the stage of development [83]. Recently, there are 18 approved and currently in-use COVID-19 vaccines [48]. The COVID-19 vaccines are generally divided into four categories: 1) Whole virus vaccines, 2) Protein-based vaccines, 3) Viral vector vaccines and 4) Nucleic acid vaccines.

As a matter of fact, even though the COVID-19 danger is decreasing a crucial role for achieving this result has been played by vaccines that still are recommended as the most important tool for the total elimination of this dangerous virus. However, since their introduction the people are debating on the safeness of this vaccines as they were released in one year while normally the release of a new drugs took five to ten years and rational arguments like the fact that now pharmaceutical technology is quite more accurate and data analysis helps in better interpretation of the experimental results find their counterpart in the argument that measuring the side effects requires years of observation. However, as for physical items it is important to consider the user preferences to find a good match also for information a proper way of presenting them could help users to take a more aware decision.

1.6 Geographic Information System and complex data

Geographic Information Systems (Geographical Information System (GIS)) are pivotal for managing and understanding complex data, particularly in spatially diverse contexts [91]. GIS enables the integration of diverse datasets, incorporating geographic elements that add a crucial layer of complexity. This technology allows for the visualization, analysis, and interpretation of complex relationships within data, offering insights into spatial patterns, trends, and correlations. In sectors such as urban planning, environmental science, and disaster management [91], GIS provides a powerful tool for decision-makers to comprehend intricate spatial data, optimize resource allocation, and formulate informed strategies. By facilitating a holistic understanding of complex data through a spatial lens, GIS enhances the efficiency and accuracy of decision-making processes across various domains.

GIS allows health authorities to analyze demographic data, population density, and healthcare infrastructure in different regions. This analysis aids in identifying areas with high-risk populations, areas with limited access to healthcare, and places where vaccine distribution centers should be established. GIS helps in identifying and targeting high-priority groups, such as frontline workers, elderly populations, and those with underlying health conditions. By understanding the distribution of vulnerable populations, vaccination campaigns can be tailored to address specific needs efficiently. GIS enables real-time tracking of vaccine distribution, administration rates, and inventory levels. This real-time monitoring helps in identifying bottlenecks and adjusting distribution strategies as needed, ensuring that vaccines reach the right places at the right time. GIS provides an effective means of visualizing complex data, such as vaccination rates, infection rates, and population density, through interactive maps and dashboards. These visualizations aid decision-makers in understanding the situation and making informed choices. GIS helps identify areas with limited access to vaccination sites, known as vaccine deserts. By understanding these gaps, authorities can establish mobile vaccination clinics or address transportation challenges to ensure equitable vaccine distribution. GIS integrates contact tracing data with geographic data, allowing authorities to identify and respond to emerging hotspots of infection. This information can guide targeted vaccination efforts to contain the spread of the virus. GIS can be used to track the effectiveness of vaccination campaigns and identify any adverse effects or side effects in specific regions. This data helps in adjusting strategies and addressing any safety concerns promptly. GIS-based maps and data visualizations can be used to inform the public about vaccination sites, availability, and eligibility criteria. This enhances public awareness and encourages people to get vaccinated.

In summary, GIS is an invaluable tool for managing COVID-19 vaccination efforts. It enables data-driven decision-making, targeted interventions, and optimized resource allocation, contributing to the successful control and containment of the pandemic through vaccination.

1.7 Thesis Structure

The rest of the thesis is structured as follows.

Chapter 1 introduces the problems and objectives of conducting research on that problem.

Chapter 2 presents the background of the problem.

Chapter 3 describes the details of sentiment analysis, wherein a hybrid approach combining BERT and NBSVM is employed.

Chapter 4 presents the recommender system for COVID-19 vaccination.

Chapter 5 explains the geo-spatial approaches useful for the control and monitoring of COVID-19.

Chapter 6 summarizes the overall thesis.

Chapter 2

Background and Related Work

In this Chapter[1], considering the case study of COVID-19 as an example of complex data, we have performed the survey of thirty primary studies related to sentimental analysis and recommender systems during COVID-19 pandemic and figure out the techniques that have been applied in order to classify the sentiments of the people as well as the application areas of sentimental analysis during COVID-19 research. The objectives of this survey are to identify the data sources and data volume of sentimental analysis during COVID-19, to identify the mostly used approaches and the applications of sentimental analysis during COVID-19. This study also presents the future implications of research with respect to COVID-19.

2.1 Methodology

The review of thirty primary studies has been conducted in this study as shown in Table 2.1. In Table 2.1, benchmark data sets and well known data sources are mentioned in column 2 to help the researchers or readers in getting similar kind of data. The volume of data used in individual study has been mentioned in column 3. The Column 4 specifies the types of approaches or techniques which have been widely used during COVID-19 for sentimental analysis and classification. During COVID-19, sentimental analysis was performed over different application areas which have been

[1] The content of this Chapter is mainly based on References 3 of the Author's publication list.

illustrated in column 5 of Table 2.1. This is the most important aspect of this surveys as it can open new research directions or topics for future researchers. The future trends and implications have been presented in column 6.

2.1.1 Data Sources during COVID-19 research:

Sentimental analysis is considered as a sentimental classification task. During COVID-19 pandemic, people experience different emotions and express their emotions using different social media platform. The social media platforms are the rich source of information as well as data in order to figure out the people's reactions and feelings during the destruction of COVID-19. Table 2.1 shows that the biggest data source for the research during pandemic was **twitter**.The statistics shows that 24 out of 30 studies uses twitter as a data source while other sources of data are online media and forums, Weibo account, WeChat account, Reddit, Yelp, RateMDs, HealthGrades, and Vitals and Qingbo Big Data Agency. The information which can be found on these popular social media is given in table 3.1.

Twitter: Twitter is considered as most popular social media platform having almost 81.47 million registered users [2]. People share message, that are called "tweets", related to public and global situations ultimately turning the twitter into data hotspot for web-based media conversation. In a single day, people post about 500 million tweets which results in 200 billion tweets posted per year [14]. The tweets are grouped based on their topics such as political matters, personal opinion, national economic issues, COVID-19 pandemic [29].

WeChat: WeChat is the Chinese multi-purpose social media and messaging platform. It has one billion monthly active users which regarded the WeChat as most popular social media platform.

2.1.2 Approaches for COVID-19 Sentimental Classification

With the rise of big data, there is a need to develop efficient analytics tools [30].Sentimental Classification Approaches, during COVID-19 research, can be divided into three types. Machine learning based approaches, lexicon based approaches and hybrid approaches.

Machine Learning Approaches:

The machine learning based approaches use the famous ML algorithms for the SC during COVID-19. They further consists of two categories i.e. supervised and unsupervised learning methods.

Supervised Learning Methods: In the supervised learning methods, the instances of the data are labelled already [81]. Various supervised learning methods have been used in literature for the sentimental classification in COVID-19 related research as seen in Table 2.1.

Naive Bayes: Naive Bayes is one of the supervised learning algorithm and have been used in [2] and [65]. It works on the principle of Bayesian theorem given in equation 2.1.

$$P(H|X) = P(X|H)P(H)/P(X) \tag{2.1}$$

Support Vector Machine: SVM is the statistical learning based machine learning algorithm that works by converting feature space into high dimensional features in order to find the hyperplane. It is used by [2], [51], [62] and [85].

Decision Tree and Random Forest: Decision tree Decision Tree (DT) is the machine learning algorithms that trains its model to predict the class values based on simple decision rules found in entire train dataset. Random Forest Random Forest (RF) belongs to the family of decision tree and works by choosing random features as well as random instances. It has been used by [2], [41], [62] and [85].

Other supervised learning approaches used in SC for COVID-19 research are K-Nearest Neighbour (KNN) [2], Linear Regression (LR) [2], Logistic Regression (LoR) [65], [85], Long-Short Term Memory (LSTM) [32], [62], RNN [50] and Bidirectional Encoder Respresentation of Trasnforemrs (BERT) model [14], [47] etc.

Unsupervised Learning Methods: In unsupervised learning methods, the data is not labelled. The unsupervised methods have been used in SC related to COVID-19. K-means clustering has been used by [24] while Latent Dirichlet Allocation (LDA) method has been used in many studies i.e. [20], [29], [56], [70], [85], [86], [93].

Lexicon Based Approaches:

Two types of opinion words are used to express the feelings i.e. positive opinion words and negative opinion words which are used to express likes and dislikes respectively. Different approaches are used to collect the opinion words list.

Dictionary-based approach and Corpus-based approach Dictionary based and corpus based approaches have been used in [5] in order to perform the sentimental analysis during COVID-19

Natural Language Processing NLP is used along with lexicon based methods in order to find the semantic relationship in a sentence. It has been used in [4] for the mental health analysis of students during COVID-19. In [50] and [58], NLP has been used to Analyze sentiments and characteristics of Covid-19 respectively. [86] has also used NLP to examine COVID-19–related discussions, concerns, and sentiments using tweets.

2.1.3 Sentimental Analysis Applications during COVID-19

COVID-19 has attracted the researchers in the area of sentimental classification as COVID-19 has effected people's behaviours and attitudes in many ways. There are various topics that work under sentimental classification during COVID-19.

Sentimental analysis on palliatives distribution during COVID-19

It is responsibility of any government to maintain the sustainability of any country. During COVID-19, people needed help in order to lessen the economic as well as psychological stress. For this purpose, different governments releases the relief packages and certain other bonuses. In developing countries, monitoring the public funds transparency is a challenge. Therefore, it is necessary for the government to analyse the people' reaction and sentiments on the palliative distribution as it will indicate the reach of funds and its impact on people's circumstances during COVID-19. In [2], Adamu et al. performed sentimental classification on on Nigerian Government COVID-19 Palliatives Distribution

Public sentiments and mental health analysis of students during the lockdown

COVID-19 has stopped the lives of people as it spreads with the human-to-human interaction. The one measure which was adopted by almost all the states of the world is "lockdown", resulting in closing airspace, closing educational institutions and workplaces, closing public transport etc. Hence, these implications have caused sadness, loneliness, anxiety, fear and many other psychological issues in peoples specially in students. Some of the students have stuck in their hostels, far away from their hometowns, few students are worried because of their exams and educational activities. Thus, the lockdown has effected people's lives and emerged many physiological issues like depressions. During these days, people are using social media in order to express their feelings and emotions. These social media posts such as tweets can be analyzed and helps the researchers to understand the state-of-mind of the citizens [20], [56], [5] and students [4]. In [24], [50], [60], [62], [65], [58], [42], [70], [86], [29], [67], [90], sentimental analysis of people's behaviour and attitudes during COVID-19 has been performed using twitter data. In [85], the sentimental analysis of tweets is carried out with respect to the age of the social media users and they found out the extent of tweets is higher in youth during COVID-19.

COVID-19 reopening sentiments:

With the paradigm shift due to COVID-19, billions of people's life has been effected directly or indirectly. COVID-19 has induced the feelings of fear, anxiety as well as economical crisis, which altogether are the challenges towards the reopening after COVID-19 [64]. Long-term lock-down is not a solution, instead a threat for the economy of any country. Considering this situation, everyone is craving for going back to normal life and physical activities [44]. Hence, in [44] and [64], the researchers tried to analyse the sentiments of the people towards reopening after COVID-19 disasters.

Analyzing online restaurant reviews

This era of e-commerce has enabled the customers to led a satisfy and quality life. The online reviews has helped the customers in decision-making. Online reviews are important for the restaurants as well because they are aligned with the star rating and one-star increase can earn a good revenue for the restaurant. The researchers analyzed the customers' sentiments which in-turn helped the customers as well as restaurants management to get good quality food and environment and maintain high quality respectively [41].

Vaccine sentiments and racial sentiments

In [47], researchers used concept drift in order to classify the sentiments of people associated with COVID-19 vaccine. During COVID-19, a rise in prejudice and discrimination behaviour against Asian citizens have been seen. The researchers tries to describe variations in people attitudes towards racism before and after COVID-19 [51].

2.2 Comparison of Studies

In this study, the comparison of thirty primary studies have been performed and represented in the tabular form in Table 2.1. The table provides an overview of various approaches and applications employed for sentiment analysis during the COVID-19 pandemic. Several data sources, including Twitter, online surveys, and review platforms like Yelp, have been utilized, with data volumes ranging from thousands to millions of entries. Techniques such as Natural Language Processing (NLP), machine learning algorithms (KNN, RF, NB, SVM, DT, LR), and deep learning methods (LSTM, BERT) have been applied for sentiment analysis. The applications cover a wide range, from analyzing public sentiments on COVID-19-related topics to mental health analysis and assessing attitudes towards the vaccine. However, limitations and future directions are also identified, including challenges related to data size, language considerations, real-time classification, and the need for broader representation and exploration across different platforms and languages. The table underscores the diverse strategies employed to understand public sentiment during the pandemic and the ongoing efforts to improve and expand sentiment analysis methodologies.

Table 2.1. Sentimental Analysis Approaches and Applications During COVID-19 Pandemic

Ref	Data Source/s	Volume of Data	Techniques	Application	Limitation/ Future Direction
[2]	Twitter	9803 Tweets	KNN, RF, NB, SVM, DT, LR	Sentimental analysis on COVID-19 Palliatives Distribution	Large data, consider multiple language, real time classification
[4]	Twitter	330,841 tweets	NLP, bar graph,	Mental Health Analysis of Students	N/A
[57]	Twitter	73,760 tweets	LDA	Attitude towards COVID-19 vaccine	N/A
[14]	twitter	3090 tweets	BERT Model	Classifying the fake tweets	N/A
[20]	Twitter	410,643 tweets	Scatter plot, line chart, LDA	Public sentiments during the lockdown	Focus on English language, understand the perceptions and contexts related to negative sentiment
[5]	Twitter	6,468,526 tweets.	Dictionary-based methodology and corpus-based methodology	Sentiment Analysis During COVID-19	N/A
[24]	Online survey	N/A	clustering algorithm (k-means)	Understand adults' thoughts and behaviors	N/A

[28]	Online media and forums, Weibo account, WeChat account	N/A	Correlation analysis	Construct a framework of COVID-19 from five Dimensions i.e. epidemic, medical, governmental, public, and media responses	N/A
[29]	Twitter	N=1,001,380	Latent Dirichlet Allocation	Sentimental Analysis, Identify dominant topics during COVID-19	Population is not represented, real-time posting
[32]	reddit	563,079 Comments	LSTM	Uncover issues related to COVID-19 from public opinions	Evaluate other social media using hybrid fuzzy deep-learning techniques
[41]	Yelp	112,412 reviews	GBDT, RF, LSTM, SWEM	Analyzing online restaurant reviews	Different review platforms and restaurant locations.
[42]	Twitter	20,325,929 tweets	CrystalFeel	Examine worldwide trends of fear, anger, sadness, and joy	Expanding the scope to include other media platforms.
[27]	Twitter	500,000 tweets	TextBlob	Determining polarity and subjectivity in COVID-19 tweets.	Explore other social media
[47]	Twitter	57.5M English	BERT	Concept drift on vaccine sentiments	Concept drift in real-time social media monitoring project

[50]	Twitter	N/A	NLP, RNN	Analyze sentiments and manifestations	Visualization, clustering and classification
[51]	Twitter	3,377,295	SVM	Changes in racial sentiment	Examine longer-term temporal changes in racial attitudes
[53]	Twitter	840,000 tweets	TextBlob, LDA,	Attitude of Indian citizens while discussing the anxiety, stress, and trauma	Analyzing how perception changes for different biographies
[58]	Twitter	57 454 tweets	NLP and text analysis	Analyse the characteristics of polish COVID-19	N/A
[60]	Twitter	370 tweets	subjectivity vs. polarity, WordCloud	Sentimental analysis for COVID-19, CORONA VIRUS, COVID–19	N/A
[44]	Twitter	293,597 tweets	Binary logit model	Understanding of the factors driving post-COVID-19 reopening sentiment	Socioeconomic and household information is averaged at the state level which provides a little variation
[62]	Twitter	7528 tweets	TextBlob, CNN-LSTM, RF, SVC, ETC, DT,	Perform COVID-19 tweets sentiment analysis using a supervised machine learning approach	Use deep learning approaches in future

[65]	Twitter	900000 tweets	Syuzhet and sentiment (R packages). NB, LR,	Public sentiment associated with the progress of Coronavirus	Include other social media platforms, news articles and personal communications data.
[64]	Twitter	293,597 tweets	N-gram, R packages Syuzhet and sentimentr	COVID-19 US Reopen Sentiment Analytics	Can be replicate on other social media data
[67]	Twitter	16 million tweets	A clustering-based classification and topics extraction model, TClustVID	Investigate Topics and Sentiment in COVID-19 Tweets	Explore other data repositories.
[70]	RateMDs, HealthGrades, and Vitals	55,612 PORs of 3430 doctors	TF–IDF, LDA	Illustration of how U.S. patients express their views during the early period of the COVID-19 crisis	Investigate these emerging trends in regions where death and recovery rates were faster.
[85]	Twitter	Twitter data (N = 82,893)	LR, SVM, RF, LDA	Examined public discourse and sentiment regarding older and COVID-19 and assessed the extent of ageism.	Twitter is popular among youngsters and the users have mostly similar attitude.

[86]	Twitter	4 million Twitter messages	LDA, NLP	To examine COVID-19–related discussions, concerns, and sentiments using tweets	Explore public trust and confidence in existing measures and policies, which are essential
[90]	Twitter	13 million tweets	Dynamic Topic Models (DTM)	Detecting Topic and Sentiment Dynamics Due to COVID-19 Pandemic	More specific topics can be analyzed to help policy maker, government and local communities during any emergency conditions.
[93]	Qingbo Big Data Agency	N/A	LDA	Social media topics and emotional change characteristics are analyzed from spatiotemporal perspectives	More precise location information can improve spatial analysis.

2.3 State-of-the-art-algorithms

Decision tree and random forest algorithms belong to the same family, and they are able to learn from user interests [89]. Random forest algorithms usually avoid some pre-defined assumptions and work well when the dataset is non-linear and contains high-order interactions. Random tree works by performing classification and regression trees and their votes. It chooses random samples as well as random features from the dataset and generates binary trees for training the model. It uses one-third of the class for testing [61].

On the other hand, decision tree generates simple decision rules from the entire dataset and trains its model on those rules to predict the class values. Decision tree algorithms are well-suited when the dataset is small [6]. Another widely used machine learning algorithm is Naive Bayes. It assigns equal weights to all features and considers them statistically independent, which means that the feature values do not exhibit any relationship [6, 1]. Naive Bayes calculates the probability of each feature using the Bayesian theorem reported in Equation 2.2:

$$P(H|X) = P(X|H)P(H)/P(X) \tag{2.2}$$

where, H: represents the sentiment class (positive or negative), X: represents the input data, such as a text document, P(H): represents the prior probability of the sentiment class, which is the probability of the sentiment class occurring without considering the input data, P(X): represents the prior probability of the input data, which is the probability of observing the input data without considering the sentiment class, P(X|H): represents the conditional probability of the input data given the sentiment class, which is the probability of observing the input data given that it belongs to a specific sentiment class, P(H|X): represents the posterior probability of the sentiment class given the input data, which is the probability of a particular sentiment class given the observed input data.

Another widely used algorithm is KNN that works by searching for the most similar instances in the whole dataset, which in turn requires a huge amount of time for processing. Hence, it should only be used with simpler and smaller datasets. The interesting fact about KNN is that it does not form a test-train model. Rather, it searches for the nearest value in the dataset. The parameter used for searching is the number of neighbors, which is provided by users [1]. KNN uses the formula reported in Equation 2.3 to search for the similar sample:

$$di = \sqrt{[(xi - x)^2 + (yi - y)^2]} \tag{2.3}$$

where, di is the Euclidean distance between a point (xi, yi) and a reference point (x, y), xi is the x-coordinate of the point being considered, yi is the y-coordinate of the point being considered, x is the x-coordinate of the reference point, y is the y-coordinate of the reference point.

Support vector machine algorithms are quite effective in high-dimensional feature space. Indeed, SVM generates a hyperplane which is used for classification. SVM produces a single feature by combining features from different sources and then trains the model. The hyperplane with the largest separation between the points of two different classes is chosen for classification. SVM leverages linear, polynomial, sigmoid, and radial basis function (RBF) as kernel functions [38].

Finally, BERT (Bidirectional Encoder Representations from Transformers) works on masked language (MLM) by using a word representation model. BERT uses separation tokens [SEP] and classification tokens [CLS] and takes the [CLS] token as the initial input, which is further enriched by word sequences. It then transfers the input to upper layers, where the self-attention mechanism is applied. The result is then directed to the upcoming encoder through the feed-forward network. The obtained vector C represents the output of the model, which can be used for multiple purposes, such as classification and translation, to name a few. The probability of sentiment classes can then be computed by equation 2.4 [77]:

$$P = softmax(CW^T) \tag{2.4}$$

where, P: is the probability distribution over sentiment classes, where each element of P represents the probability of the input belonging to a particular sentiment class, softmax(): is a function that maps a vector of arbitrary real values to a probability distribution such that the output values are non-negative and sum up to 1, C: is a matrix that contains the learned representation of the input data, also known as the embedding matrix, W: is a matrix of learned weights that map the input representation to the sentiment class probabilities, T: is the transpose of the weight matrix W.

Chapter 3

Improving AutoEncoder for Sentiments Classification

In this chapter[1], examining the COVID-19 case study within the context of complex data, we used twitter (now X app) complex data to analyze people's response and concerns about COVID-19 vaccination all over the world using sentiments analysis. We proposed a model which uses X app (former twitter) data and perform necessary pre-processing steps for removing stop-words, hashtags and URLs. The analysis further finds the polarity of tweets and make word clouds of positive tweets, negative tweets and neutral tweets and then performed analysis of people's sentimental using high performance model.

3.1 Research Questions

1. RQ1: What are the key concerns and sentiments expressed by individuals in the analyzed X app data regarding COVID-19 vaccination?

2. RQ2: How effective is the proposed sentiment analysis model in classifying and visualizing sentiments from X app (former Twitter) data?

[1] The content of this Chapter is mainly based on References [76], [80] and [79] of the reference list.

3.2 Methodology

In this chapter, we propose a hybrid approach for sentiment analysis.

3.2.1 Proposed Framework

In order to properly address our sentiment analysis task related to COVID-19 tweets, we propose a methodology composed of the sub-tasks, as shown in Figure 3.1.

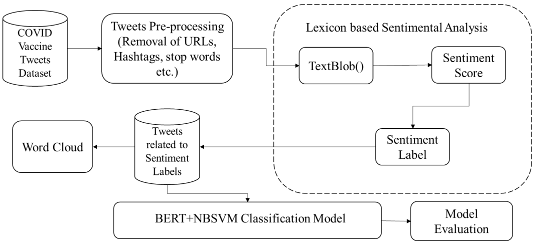

COVID Vaccine Tweets Dataset
Tweets Pre-processing (Removal of URLs, Hashtags, stop words etc.)
Lexicon based Sentimental Analysis
TextBlob()
Sentiment Score
Sentiment Label
Tweets related to Sentiment Labels
Word Cloud
BERT+NBSVM Classification Model
Model Evaluation

Figure 3.1. The overall proposed methodology of our research

First, we deal with dataset collection and pre-processing. Indeed, tweets are collected from their sources in the original unstructured format, thus we need to extract main features for subsequent analysis and eliminate unnecessary and noisy information like stop words and Uniform Resource Locators (URL) to cite a few.

Sentiment classification is then performed by extracting the sentiments and their polarity value using lexicon based approaches. As polarity values are computed, we perform sentiment classification by our innovative BERT+NBSVM model.

3.2.2 Collection of data

In this experimental study, we analyzed a substantial dataset comprising 70,000 instances of data. Each instance represents a unique entry, such as a tweet or record, collected from X app related to COVID-19. This extensive dataset allows for a comprehensive exploration of sentiments of people, providing valuable insights into COVID-19 duration. The large volume ofsc instances enhances the robustness and generalizability of our findings, offering a thorough understanding of peoples' sentiments during COVID-19.

We used Twitter API through the Tweepy Python package to collect tweets containing the terms Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford/Astra-Zeneca, Covaxin, and Sputnik V. The attributes we extracted from the dataset and their description are reported below:

- ID: Unique identifier for each tweet
- user_name: Screen name or username of the Twitter account that posted the tweet
- user_location: Location listed on the user's Twitter profile
- user_description: Bio or description listed on the user's Twitter profile
- user_created: Date the user's Twitter account was created
- user_followers: Number of followers for the user's Twitter account
- user_friends: Number of accounts the user is following on Twitter
- user_favourites: Number of tweets the user has favorited on Twitter
- user_verified: Boolean indicating if the user's Twitter account is verified
- date: Date and time the tweet was posted
- text: Content of the tweet
- hashtags: Any hashtags included in the tweet
- source: The device or application used to post the tweet
- retweets: Number of times the tweet has been retweeted
- favorites: Number of times the tweet has been favorited
- is_retweet: Boolean indicating if the tweet is a retweet or an original post.

Oversampling of Imbalance Dataset In classification scenarios, encountering an uneven distribution of target class labels is a common issue. This condition, known as an imbalanced dataset, has a notable impact on the training process of data mining models. Primarily, the model tends to focus on the majority class during training, leading to biased class predictions. This bias arises from the fact that the minority class, containing fewer instances, may be treated as noise or outliers. Consequently, addressing imbalances in the data is crucial and should be considered as an essential preliminary step before conducting classification. We used BORDERLINE-SMOTE to perfrom the oversampling of imbalance data. Termed as borderline-SMOTE, this variant of SMOTE operates with a distinctive approach. Instead of generating synthetic data randomly in proximity to existing data, borderline-SMOTE specifically identifies the borderline for each class. Instances positioned on the borderline and in close proximity are considered more prone to misclassification compared to those situated farther away. Consequently, these instances play a more crucial role in the classification task. In the context of borderline-SMOTE, all instances in the minority class are categorized into three groups: noise, which is rare, incorrectly positioned in areas predominantly occupied by the majority class; danger instances, located on class boundaries and overlapping with the majority class; and safe instances, representing the minority class.

Data Pre-processing and Noise Removal. Noise is a factor that affects the analysis badly. It was observed that collected data had a lot of noise i.e. special characters, punctuations, numbers and emojis. The presence of noise badly impact the quality of classification results As it is easy to see we selected only relevant attributes for our analysis. To this end, the dataset has been cleaned by removing URLs from the text, removing the hashtag symbol from the tweets and finally removing stopwords. To perform this task we wrote ad-hoc Python scripts. In Table 3.1, we show some examples of the results obtained after tweets pre-processing.

3.2.3 Getting Sentiment Polarity Values

In order to perform a valuable sentiment analysis, polarity values computation for tweets is crucial. Polarity indicates whether a given sentence falls in positive category or negative category based on some pre-defined

Table 3.1. Comparison of tweets before and after pre-processing

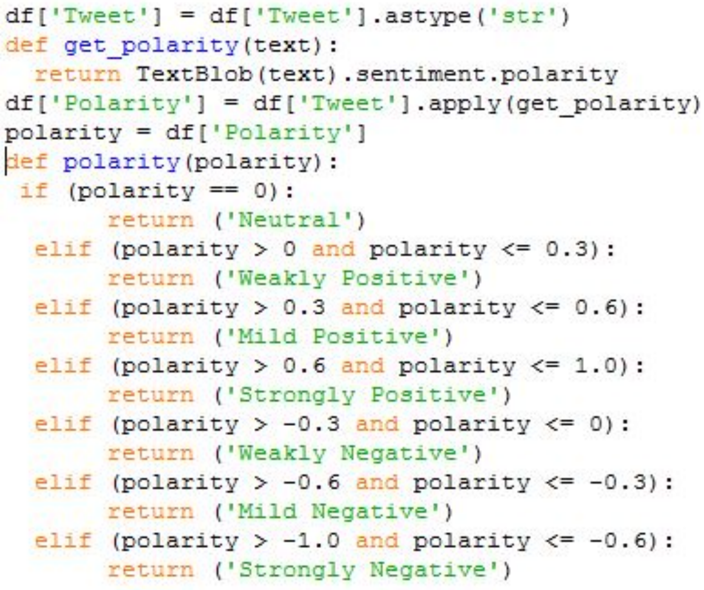
Dummy samples	Hashtags removal	URLs removal
Fever after first dose #PfizerBioNTech https://t.co/xffiee77	Fever after first dose PfizerBioNTech https://t.co/xffiee77	Fever after first dose PfizerBioN-Tech
Vaccine scheduling available online https://t.co/jgeeityc	Vaccine scheduling available online https://t.co/jgeeityc	Vaccine scheduling available online
Second dose done?? https://t.co/ooehdugy	Second dose done https://t.co/ooehdugy	Second dose done

classification. We leveraged the classification proposed in [73] and categorized our tweets into seven sentiment classes based on the values of their sentiment polarity. The sentiment classes are defined as: *neutral, weakly positive, mild positive, strongly positive, weakly negative, mild negative and strongly negative.* We fixed polarity range of each class using principles of [73] as shown in figure 3.2.

We find the polarity values (between [-1 to +1]) of the given tweets using the TextBlob() library function of Python. The working principle of TextBlob() can be seen in Figure 3.3.

3.2.4 Word Clouds

The polarity values are then used to compute the overall classification of the tweet as positive, negative, or neutral. This information about the positive, negative, or neutral tweet classification can be used to generate word clouds. In detail, each word cloud assigns word sizes according to their occurrence frequencies. In a word cloud, a bigger word size denotes a higher number of occurrences of the word. We plotted the word cloud using Python scripts to observe the high-frequency words in positive, negative, and neutral clouds. This step is particularly useful since the most frequent words can help identify the dominant opinions and feelings of people regarding a specific vaccine issue.



```
df['Tweet'] = df['Tweet'].astype('str')
def get_polarity(text):
  return TextBlob(text).sentiment.polarity
df['Polarity'] = df['Tweet'].apply(get_polarity)
polarity = df['Polarity']
def polarity(polarity):
 if (polarity == 0):
       return ('Neutral')
  elif (polarity > 0 and polarity <= 0.3):
       return ('Weakly Positive')
  elif (polarity > 0.3 and polarity <= 0.6):
       return ('Mild Positive')
  elif (polarity > 0.6 and polarity <= 1.0):
       return ('Strongly Positive')
  elif (polarity > -0.3 and polarity <= 0):
       return ('Weakly Negative')
  elif (polarity > -0.6 and polarity <= -0.3):
       return ('Mild Negative')
  elif (polarity > -1.0 and polarity <= -0.6):
       return ('Strongly Negative')
```

Figure 3.2. Threshold of polarity for sentiment classes.

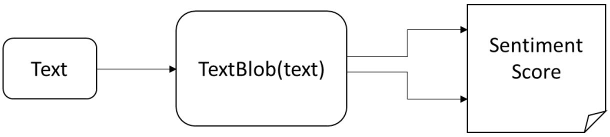


Figure 3.3. How TextBlob() works?

3.2.5 Generic BERT for Sentimental Classification

For the sentimental classification of the tweets related to COVID-19 vaccines, we used high performance approaches such as BERT model in our work. BERT, Bi-directional Encoding Representation for a Transformer, model is works on masked language (Maked Language modelling (MLM)) by using the word representation model. The bidirectional transformer is used for its training. In past, only two language models were available that were unidirectional, such as right to left and left to right. However, the BERT used MLM and hence has to ability to predict random masked words in the sentence. So, it can be used for learning bidirectional representation. In natural language, using representation from bidirectional instead of unidirectional is very crucial. BERT is one of the most famous model which has architecture of modern language [55] [87]. BERT base model and BERT large model are the two main architectures of BERT. There are four major difference in both of the models, that are hidden layers of encoder, the number of self-attention, feed-forward network hidden size and the maximum sequence length parameter [55]. We used face and encoder for BERT implementation and performed sentiment classification of vaccination tweets using these steps.

- We divided the data in two sets i.e. test and train
- We used training set for representation of torch tensors
- We used batch size for tensors and iterators for fine tuning the model
- We trained our model and validated its accuracy
- We performed the model evaluation.

BERT System Architecture

In BERT system architecture, fine tuning is performed on the outermost layer. The training of the core architecture is done using text corpora and the internal layer is frozen. The BERT architecture BERT can be seen in figure 3.4

BERT uses [Seperation (SEP)] and [Classification (CLS)] as separation and classification tokens. The BERT takes [CLS] toekn as initial input which is further accomplished by words sequences as input. It then transfers the input to the upper layers, where self-attention is applied. The output of which then referred to the feed-forward network, and forwarded to up coming encoder. The vector C is the output of the model. This vector can be used for multiple purposes, such as classification and translation etc. The probability of sentimental classes can be calculated by following equation 2.4.

Transformer

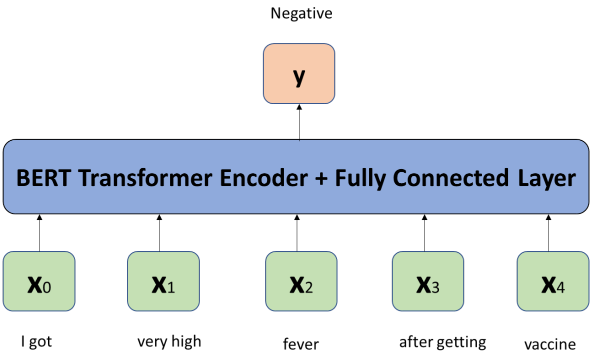


Figure 3.4. Architecture of BERT model

In transformer, each encoder consists of six layers. These identical six layers have two further layers such as feed forward network layer and multihead self attention layer [9] [21]. In literature, few studies used residual connections and normalization for each sub-layer as well. Similarly, the decoder also has same identical layers, but the layers further consists of three sub-layers. The additional sub-layer in decoder is found on the output of the encoder and called multi-head self-attention layer [22]. In decoder also, the layer normalization and residual connections have been provided [13]. In decoder stack, the multihead layer ensures that the position less than i is responsible for the prediction of position i [25].

The attention has three components such as the queries, the keys and the values, which are denoted as Q, K and V respectively [45]. In this article, we used the queries and keys of d$_k$ dimensions and values of d$_v$ dimensions [39]. We computed the dot product of the keys and queries [72]. The weights are obtained using SoftMax function [25]. The following formula was used to calculate the attention in equation 3.1.

$$Attention(Q, K, V) = SoftMax(\frac{QKT}{\sqrt{dv}})V \tag{3.1}$$

Where matric Q represents the set of queries, and set of keys are packed into K and values are packed into V.

In multi-head attention, projections of values, queries and keys are projected h times to d$_k$, d$_k$ and d$_h$. Then the attention function is performed parallel over theses projected values, queries and keys. Then, they are concatenated and again projected to gain the final output.

Multi-Head (Q, K, V) = Concat (head_1; :::; head_h)W^O

where head_i = Attention ($QW_i{}^Q$, $KW_i{}^K$, $VW_i{}^K$)

There are three ways to use multi-head attention in transformers. In encoder-decoder layer, the previous decoder produce the queries and output of encoder produce values and keys. In encoder self-attention layer, the output of previous encoder layer results in all keys, values and queries and in the decoders' self-attention layer.

3.2.6 Combining BERT and Naive Bayes-SVM for Sentimental Classification

In this section, we propose a combination of BERT (Bidirectional Encoder Representations from Transformers) and Hybrid of Naive Bayes and Support Vector Machine (NBSVM) (a hybrid of Naive Bayes and Support Vector Machine) for sentiment classification of vaccine-related tweets. BERT is a well-established tool for transformers and attention mechanism implementation. The transformer is a sequence-to-sequence model based on attention mechanisms for encoding and decoding textual information. However, the BERT architecture does not properly utilize the decoder potential as it only leverages the encoder layer of the transformer [55, 87]. BERT has two main architectures: BERT Base and BERT Large, which exhibit some major differences in text modeling with respect to four main features, i.e., the number of hidden layers in the encoder, the number of self-attention heads, the hidden size of the feed-forward network, and the maximum sequence length parameter [55].

On the other hand, Naive Bayes is a machine learning algorithm that performs well for short text sentiment analysis, while SVM is more appropriate for longer text. We chose to implement a hybrid approach using both Naive Bayes and SVM to obtain higher accuracy by incorporating the log count ratio obtained by Naive Bayes as a feature value in SVM [84]. This choice has been proven to be quite versatile for various analysis tasks and types of data collections.

Leveraging BERT and NB-SVM synergies.

Our BERT+NB-SVM based architecture takes advantage of the regression fine-tuned sequence-pair obtained by BERT and leverages the Naive Bayes-Support Vector Machine (NB-SVM) model [84] to obtain document-term matrices (Document Term Matrix (DTM)) that compute the Naive Bayes Log-count ratios. The latter model determines the probability that a given word appears in the document in positive versus negative classes. In a sense, we combine the strengths of deep learning and classical machine learning approaches to obtain a more accurate sentiment analysis. The system architecture of our BERT+NB-SVM based approach is depicted in Figure 3.5.

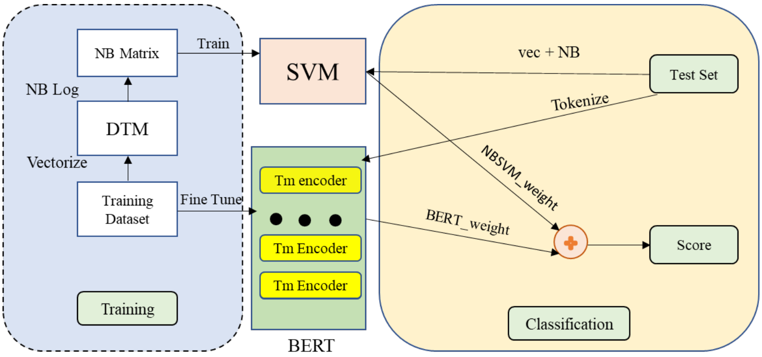


Figure 3.5. System architecture of BERT+ NBSVM

Herein, we perform the following steps for training and classification:

1. We fine-tune the BERT model on the training dataset.

2. We train an SVM model using the log count ratios obtained by Naive Bayes.

3. The final score is computed as the weighted sum of the obtained NB-SVM model and the best fine-tuned BERT model (i.e., the BERT model that exhibits the best performance over different epochs and with different batch sizes).

Hyper-parameters used for Model Training

We performed both pre-training and fine-tuning to develop our model. Specifically, we used the Adam optimizer as the loss function for training the model. Then, we performed grid search for parameter tuning, and found the best weight for the BERT model to be 0.87, while for NB-SVM it was 0.08. The table 3.2 shows the hyper-parameter that have used are in the experiments.

Parameters	**Value**
Batch Size	16
Learning rate of model	1e-5
Learning rate of Classifier	1e-3
Epochs	5
Warm up steps	0
Gradient accumulation steps	1
Max grad norm	1.0
SEED Speed	42
No Cuda	False

Table 3.2. Parameters used for training the model

State-of-the-art

To evaluate and compare the results of our model, we designed an experimental comparison with the main state-of-the-art algorithms. We performed the comparison with KNN (K-nearest neighbor) algorithm, SVM (Support Vector Machine) algorithm, RF (Random Forest) algorithm, NB (Naive Bayes) algorithm, and DT (Decision Tree) algorithm because of their wide adoption [2, 32, 62].

3.3 Achieving High Performances: The Sigma Architecture

In this section, a new architecture named *Sigma* [11] is used to provide a solution for building a complete, interactive, and scalable Big Data System using a variety of tools and techniques that achieve high-performance execution of the framework defined so far for Sentiment and Geo-Spatial analysis. After testing our framework for limited-sized datasets, we tested it in a high-performance environment to make it suitable for real-life scenarios where data size quickly increases, thus requiring a proper architecture to deal with it. The typical working scenario collects data from real-time sources that need to be collected and analyzed properly, as they could quickly exceed common computational facilities, in order to make them suitable for the analysis steps. In order to address the aforementioned computational issues, we describe in this section the Sigma Architecture that differs from well-known Lambda [82] and Kappa [37] architectures, which are considered the reference architectures for supporting the tasks described in our framework.

Sigma Architecture is composed of three Layer as depicted in Figure 3.6.

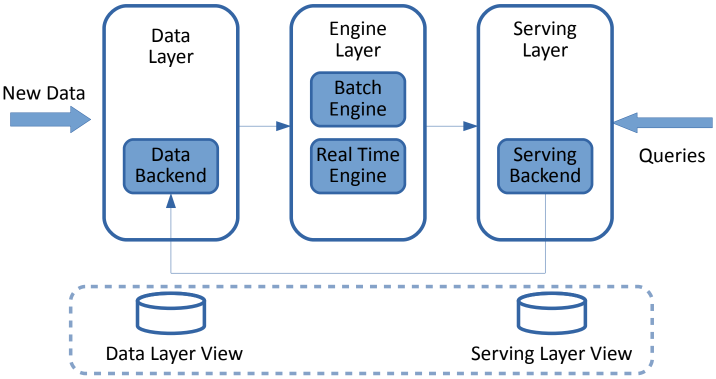


Figure 3.6. Sigma Architecture

The **Data Layer** stores a copy of all raw data that are collected by the system. This layer stores an immutable, constantly growing dataset (*Data Layer View*) and offer a back-end able to perform random reads on the whole content. The **Engine Layer** is responsible to compute arbitrary function on the data layer and to store results on the Serving Layer. This computation can be executed in batch mode on the whole Data Layer View (via Batch Engine) or can be executed in real time mode every time new data arrives in the Data Layer (via Real Time Engine). The **Serving Layer** is a specialized distributed database that loads in the results of Engine Layer Computation (*Serving Layer View*) and makes it possible to do random reads on it.

3.4 Results and discussion

In this section, we will discuss the results we obtained by using the Sigma architecture in a Persistent Data setting as we collected the data in batches[2]. Our setup provides a data collection service with the possibility of scaling out the computing cluster to manage the increase in the volume of data to be processed. The solution includes backup functionalities to manage the database saving even when the volumes managed grow suddenly. Overall, the computing cluster provides 8 TB of storage space, 8 vcores, and 120 GB of RAM. The execution service is guaranteed by 4 virtual machine instances having the following characteristics: 4 virtual cores with Xeon Broadwell processors, 30 GB of RAM, and 100 GB of SSD persistent storage.

3.4.1 Evaluation Matrices

We trained one model for positive tweet classification, and a second one for the classification of negative tweets. To evaluate our model, we computed precision, recall, and F_1 score, which are defined in equations 3.2, 3.3, and 3.4, respectively:

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

$$FMeasure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3.4}$$

where: i) True Positive (TP) is true positive count and computes the number of positive prediction of positive instances, ii) True Negative (TN)

[2] Streaming data was not accessible for our experimental assessment

is true negative count and computes the number of positive prediction of negative instances, iii) False Positive (FP) is false positive count and computes the number of negative prediction of positive instances and iv) False Negative (FN) is false negative count and computes the number of negative prediction of negative instances.

3.4.2 Sentiment Polarity Computing

The sentiment of a text can be determined by its polarity value. Table 3.3 displays the polarity values and their corresponding category types for each sentence, as defined in Section 3.2.3.

Table 3.3. Polarity values and sentiment categories with respect to sample tweets

Tweet Sample	**Polarity**	**Sentiment**
Fever after first dose PfizerBioN-Tech	-0.5	Mild Negative
Vaccine scheduling available online	0.7	Strongly Positive
Second dose done	0	Neutral

In order to validate our approach, the sentiment polarity was assessed by linguistic experts who validated our assignments through a questionnaire administered to a cohort of one hundred volunteers, uniformly distributed across an 18-70 age group.

3.4.3 RQ1: What are the key concerns and sentiments expressed by individuals in the analyzed X app data regarding COVID-19 vaccination?

To accomplish our task, we computed the word clouds of the positive, negative, and neutral tweets. For this purpose, we first classified the data into three subsets using the polarity values of the tweets. Figure 3.7 displays three different word clouds for positive, neutral, and negative words. Some categories, along with their respective words, are given below:

- Words like "great", "Good", "More", "Safe", "Thank", "Better", "Happy", "Love" may indicate that people are willing to get vaccinated [3].

- Words like "Sick", "Fever", "Risk", "Hard", "Bad", "Out", "Serious" show that people are not happy with the vaccination campaign.

- Few words like "COVID", "Shot", "One", "Second", "People", "Country", "Pfizer/BioNtech", "Astrazeneca", "Sinovac", "Sinopharm", "Moderna" do not show any emotion related to vaccines.

- Words like "Receive", "Batch", "Dose", "Russia", "China", "Vaccine", "EU", "Take", are neither positive nor negative.

- Few words like "Alone", "Long", "Fail", "Still" show that people are feeling anxious during vaccination.

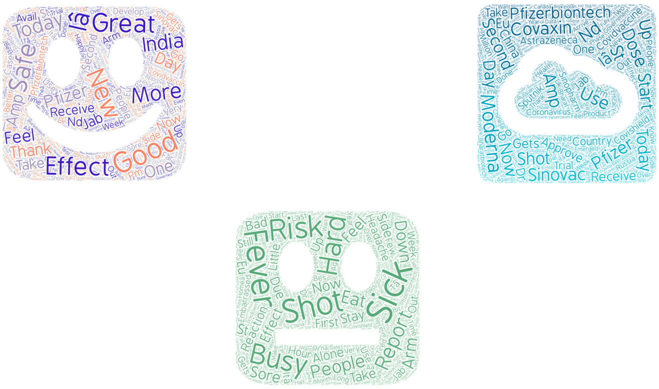


Figure 3.7. Figure contains Positive word cloud of tweets, Negative word cloud of tweets and the neutral word cloud of tweets

[3] This assumption may be controversial as the context in which these words are used plays a crucial role. Indeed, in our dataset, we leveraged some statistical analyses that suggest that, in our case, these terms may indicate a positive attitude towards getting vaccinated

3.4.4 RQ2: How effective is the proposed sentiment analysis model in classifying and visualizing sentiments from X app (former Twitter) data?

The results of the experiments are reported in Figure 3.8 and 3.9:

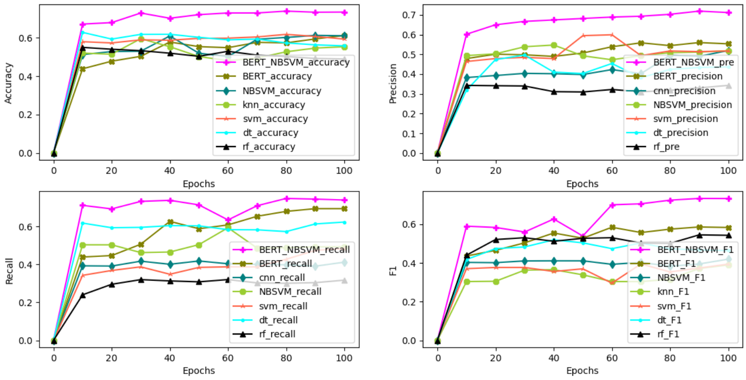


Figure 3.8. Comparison of BERT+NBSVM model with state-of-the-art for positive sentiment classification.

Figure 3.8 shows sub-graphs depicting the classification accuracy, precision, recall, and F1 score of our BERT+NBSVM model compared to BERT, NBSVM, decision trees, KNN, random forest, and SVM for the classification of positive sentiments. The results show that our approach outperformed all other state-of-the-art models by achieving the highest accuracy. [4]

Figure 3.9 shows the plots of Accuracy, Precision, Recall and F1 score for our] model compared to BERT, NBSVM, Decision tree, KNN, random forest and SVM for the classification of negative sentiments. Again our BERT+NBSVM model exhibits best performance among all other state of the art neural network and machine learning models we implemented.

[4] We would like to emphasize that we performed our comparison under the same race conditions for all approaches, using the same dataset. Therefore, our superior performance can be attributed to the fact that our hybrid model refines the results obtained by classical algorithms, improving intermediate results computed by BERT and SVM.

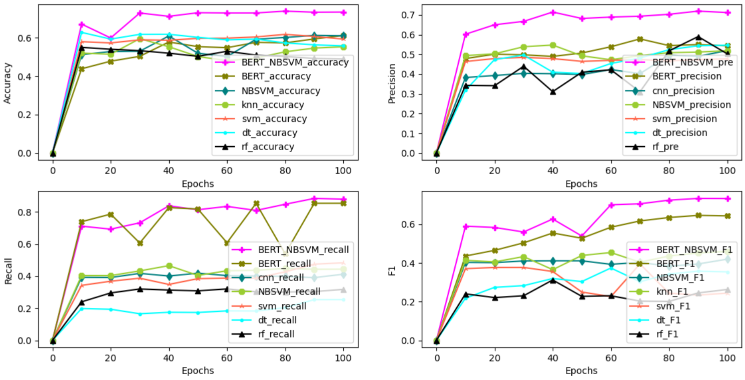


Figure 3.9. Comparison of BERT+NBSVM model with state-of-the-art for negative sentiment classification.

To wrap up, the qualitative factors that lead to better BERT+NBSVM performances can be summarized as follows:

- Feature engineering computation: NBSVM is able to extract relevant information from the text using feature engineering, thus BERT shows better feature engineering when equipped with NBSVM.

- Complementary strengths: NBSVM is well-suited for handling large datasets while BERT is well-suited for identifying the semantics of texts. Combining these models results in a synergy.

- Transfer learning: BERT is pre-trained on large texts and then fine-tuned on a specific task. Pre-training is a form of transfer learning, and when combined with NBSVM, it achieves good performance.

When considering classical machine learning models, the performance of SVM is higher than other baseline algorithms because SVM does not show any side-effects of hyper-parameters related to the data [68]. KNN and decision trees show similar accuracy, and they have a significant impact on sentiment classification [26]. Finally, Random Forest shows intermediate performance in both scenarios of our experimental setting because

Random Forest draws observation strategies randomly and requires hyper-parameter tuning to achieve better performance [59].

Chapter 4

Recommendation System for COVID-19 Vaccines

In this Chapter[1], we designed a recommendation system for COVID-19 vaccine using social media complex data as we are considering COVID-19 as a case study of complex data in this thesis.

4.1 Research Questions

1. RQ1: How does the proposed ensemble approach, combining CT-BERT-CONVLayer_Fusion and Random Forest, contribute to accurate sentiment classification of COVID-19-related tweets?
2. RQ2: To what extent does the sentiment analysis contribute to the categorization of COVID-19 vaccine-related tweets into specific predefined categories, and how does this enhance the recommendation system's comprehensiveness?

4.2 Proposed Recommendation System

The proposed architecture for the COVID-19 vaccine recommendation system, as depicted in Figure 4.1, involves several key steps. Firstly, we

[1] The content of this Chapter is mainly based on References 7 of the Author's publication list.

used the pre-processed tweets from section 3.2.2 and sentiment polarity of the tweets from section 3.2.3. Next, sentimental classification is performed using an ensemble approach consisting of CT-BERT-CONVLayer_Fusion and Random Forest. This classification process assigns each tweet to one of seven predefined sentiment categories based on its content and expressed sentiment. In the third step, similarity values between the tweets and a predefined index set of categories are calculated. These similarity values help categorize the tweets into specific categories based on their resemblance to the predefined categories. This categorization enhances the understanding of the content and sentiment of the tweets. Overall, this architecture combines tweet collection, pre-processing, sentimental classification using an ensemble of CT-BERT-CONVLayer_Fusion and Random Forest, and tweet categorization based on similarity values to create a comprehensive recommendation system for COVID-19 vaccines.

In the proposed recommendation system, when a user generates a query, a search is performed based on the preferred query keyword, such as "number of doses." The search results can then be sorted based on other categories such as country and costs to refine the recommendations. Next, the polarities (sentiments) of the reviews and vaccine pairs in the output lists are examined. This step helps determine the sentiment associated with each recommendation. Finally, the recommendation system produces an output that includes the most suitable vaccine along with reviews that align with the user's query, taking into consideration the polarity (sentiment) of the reviews. This ensures that the recommended vaccine not only matches the user's requirements but also aligns with the sentiment expressed in the reviews.

4.2.1 Sentiment Analysis

To ensure accurate sentiment analysis, a proposed model for sentiment classification employs an ensemble approach combining Random Forest classifier with CT-BERT_CONVLayerFusion model. The architecture of this sentiment classification model is illustrated in Figure 4.2.

The model combines the strengths of CT-BERT_CONVLayerFusion, which is a fusion model incorporating BERT (Bidirectional Encoder Representations from Transformers) and convolutional layers, with the Random Forest classifier. This ensemble approach leverages the deep learning capabilities of CT-BERT_CONVLayerFusion and the ensemble learning capabilities of Random Forest to improve the accuracy and robustness of sentiment classification. By combining the features and capabilities of these two models, the proposed sentiment classification model aims to provide more accurate and reliable sentiment analysis results, enhancing the understanding and interpretation of sentiments expressed in the collected data.

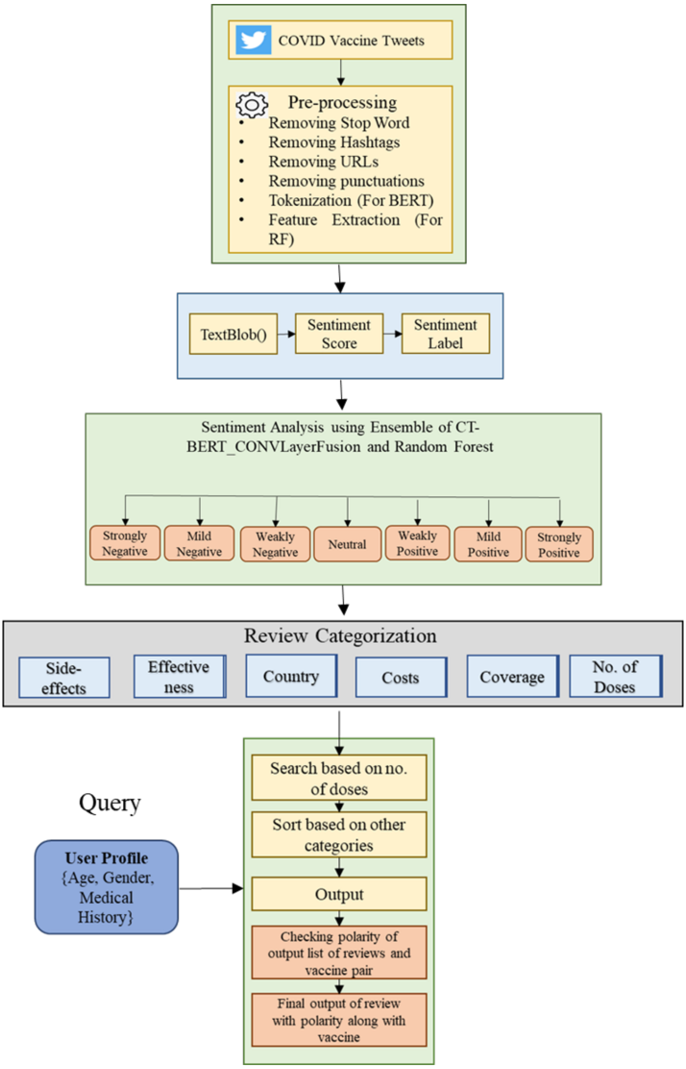


Figure 4.1. Architecture of proposed recommendation system for COVID-19 vaccines

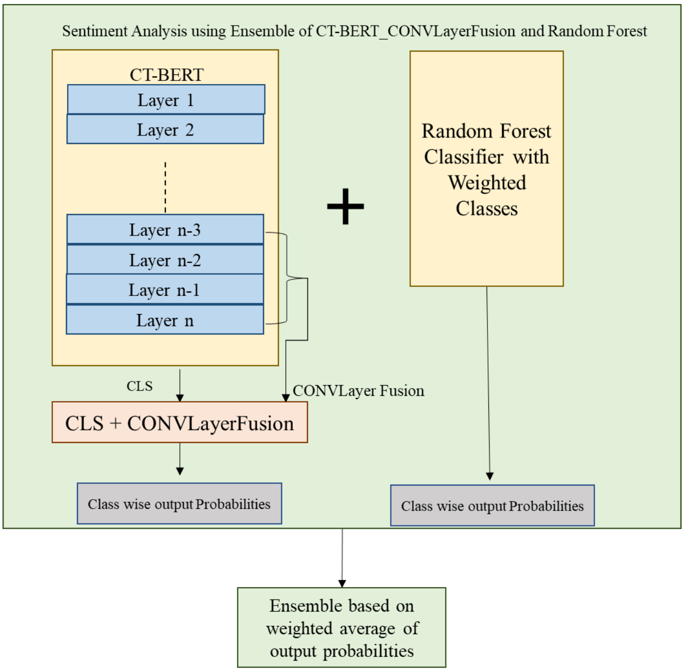

Sentiment Analysis using Ensemble of CT-BERT_CONVLayerFusion and Random Forest
CT-BERT
Layer 1
Layer 2
Layer n-3
Layer n-2
Layer n-1
Layer n
+
Random Forest Classifier with Weighted Classes
CLS
CONVLayer Fusion
CLS + CONVLayerFusion
Class wise output Probabilities
Class wise output Probabilities
Ensemble based on weighted average of output probabilities

Figure 4.2. Architecture for the proposed ensemble of CT-BERT-CONVLayerFusion with a Random Forest classifier

4.2.2 CT-BERT (COVID Twitter BERT) Model

The CT-BERT model, as described in the study by Muller et al. (2020) [47], is a transformer-based model that has been pre-trained on a large corpus of approximately 160 million tweets related to COVID-19 or coronavirus. The data used for pre-training was collected from the Crowdbreaks platform, which utilizes the Twitter API to gather English-language tweets. The collection period for the tweets spanned from January 12, 2020, to April 16, 2020.

Before training the model, the dataset underwent normalization procedures, including the replacement of usernames, URLs, and Unicode emoticons with text tokens and ASCII representation using Python. Additionally, duplicate tweets were removed from the dataset, resulting in a training corpus of 22.5 million tweets. The tweets were tokenized, and a vocabulary of 30,000 words was established. The CT-BERT model was trained using a single TPU (Tensor Processing Unit) for a duration of 120 hours. The performance of the model was evaluated by comparing it to state-of-the-art models such as BERT base and BERT large using five different datasets. The CT-BERT model demonstrated a significant improvement of 10-30% over the existing models across these datasets. The versatility of the CT-BERT model extends beyond sentiment analysis. It can be employed for various natural language processing (NLP) tasks, including text verification, identifying informative tweets related to COVID-19, and potentially other related tasks as well [12].

CONV Layer Fusion

To extract hidden sentiments in text and enhance the results obtained from pre-trained models, the proposed CONV Layer Fusion model is introduced. This model leverages the hidden layers of deep models to refine the sentiment analysis process. The architecture of the CONV Layer Fusion model is illustrated in Figure 4.3.

The CONV Layer Fusion model combines the power of convolutional layers with the hidden representations of deep models. By incorporating convolutional layers into the sentiment analysis pipeline, the model is able to capture local patterns and dependencies within the text, which can provide further insights into the hidden sentiments expressed. The fusion of convolutional layers with the hidden layers of deep models enables a more comprehensive and nuanced analysis of the sentiment within the text. This approach enhances the accuracy and depth of the sentiment analysis, resulting in improved results compared to using pre-trained models alone. Through the CONV Layer Fusion model, the research aims to uncover and capture the intricate nuances of sentiment within the text, contributing to a more refined and detailed understanding of the sentiments expressed in the data.

To enhance the CT-BERT model, the last four layers were improved by incorporating convolutional layers. This modification involved applying convolutional layers to these layers, enabling the extraction of local patterns and features. Following the convolutional layers, the MAX Pooling function was applied individually on each layer. This function selects the maximum value from each embedding dimension, reducing the dimensionality of the convolution output. After applying maximum pooling, the resulting embeddings from each layer were stacked using the PyTorch stack function. These stacked embeddings were then summed using the PyTorch sum function. This summation process combined the information from multiple layers, resulting in a refined representation of the text data.

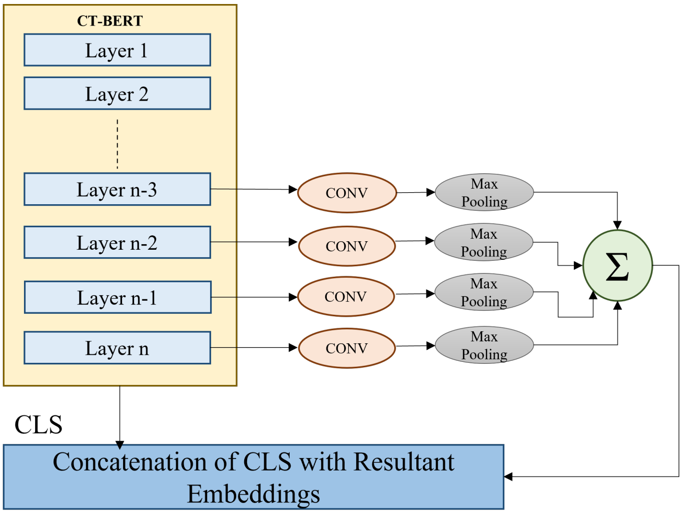


Figure 4.3. Proposed CONV Layer Fusion

The obtained embeddings were concatenated with the classification (CLS) token, which is a special token representing the entire input sequence, and then passed through a classifier to obtain class-wise probabilities. This classification step generated probabilities for different sentiment classes, allowing for sentiment analysis and classification of the input text. By incorporating convolutional layers, max pooling, and stacking of embeddings, the improved CT-BERT model can capture and leverage local patterns and features within the text, enhancing the sentiment analysis process and providing more accurate class predictions for sentiment classification.

Inserting Convolutional Layer in Last four CT-BERT layers:

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.

Furthermore, it was important to ensure that the output dimensions of the convolutional layer matched the input dimensions of the subsequent transformer layer. This compatibility allowed for seamless integration of the convolutional layer outputs into the subsequent layers of the model, preserving the flow of information and facilitating the overall sentiment analysis process.

The proposed Algorithm 1 is shown in the following.

Algorithm 1 CT-BERT-LayerFusion

```
1: Strongly Negative Tweets
2: Mild Negative Tweets
3: Weakly Negative Tweets
4: Neutral Tweets
5: Weakly Positive Tweets
6: Mild Positive Tweets
7: Strongly Positive Tweets
Conv(): Convolution layer
CT-BERT_Max(): Maximum value embeddings of each layer of CT-BERT
CT-BERT_MaxLayersSum(): Sum of Pooled embeddings
D: Dataset
Input D
Steps:
n= CT-BERT_Layers
for i=n-3 to n do
    Conv()
    Conv_Max()
end for
for i = 1 to size of (D)  do
    Final_Embeddings $\leftarrow$ Concatenation(CLS($D_k$), CT-BERT_MaxLayersSum($D_k$)
    Tweet_Classification $\leftarrow$ Classifier(Final_Embeddings)
end for
Output: Tweet_Classification (Class wise probabilities)
```

To match the input dimensions of the convolutional layer to the output dimensions of the corresponding transformer layer, we need checked the shape of the output tensor from the transformer layer and configured the convolutional layer accordingly. Here's an algorithm 2 we used to accomplish that:

Algorithm 2 Matching Input Dimensions for Convolutional Layer

```
procedure MatchDimensions(TransformerOutput)
    $[batch\_size, sequence\_length, hidden\_size]$ ← Shape TransformerOutput
    Configure Convolutional Layer:
        Set number of filters, kernel size, stride, padding
        Set number of input channels (1 for first layer, otherwise as per previous layer)
    Adjust Input Dimensions:
        Set number of channels as $hidden\_size$
        Set width as $sequence\_length$
    Adjust Output Dimensions:
        Set number of channels as number of filters
        Set width based on kernel size, stride, padding
end procedure
```

The rationale for using last four layers for improvement of the analysis is that the hidden and deeper layers contain more related information [33]. Hence, extracting information from the deeper layers and then summation of their maximum valued embeddings can produce richer information of sentiments.

4.2.3 Random Forest

Random Forest is a popular machine learning algorithm used for classification tasks. It operates by creating an ensemble of decision trees, where each tree is trained on a subset of the data and a subset of the available features. During the training phase, the algorithm randomly selects instances and features from the dataset, and this process is repeated to build a forest of trees. When making predictions, the individual outputs of the trees are combined, typically through voting or averaging, to produce a final prediction. Random Forest is known for its accuracy and robustness, and it can effectively handle high-dimensional and large datasets. In the context of sentiment analysis, feature extraction plays a crucial role in Random Forest. After performing the necessary pre-processing steps on the text data, features need to be extracted to represent the data numerically. Several common approaches for feature identification in sentiment analysis include bag-of-words, TF-IDF, and word embeddings. In the case of word embeddings, the Doc2Vec model can be utilized to convert the text data into numerical vectors. The model is trained on the available review data, learning representations of words and documents. These learned representations, known as word vectors, can be applied to the reviews of interest, generating representation vectors that serve as features for the Random Forest classification model. By employing feature extraction techniques like word embeddings, Random Forest can effectively analyze and classify text data, providing valuable insights and predictions in sentiment analysis tasks.

Table 4.1 shows the top 20 features with the highest importance, as determined by the Random Forest classifier, and these features are used for the classification task.

4.2.4 Tweets Categorization

In the categorization process, tweets or reviews are grouped into predefined categories based on the most frequent words found in the dataset. These categories represent clusters of words that commonly appear together in the dataset. Figure 4.4 illustrates the visualization of these categories.

By identifying the most frequent words and grouping them into categories, we gain insights into the main themes or topics present in the dataset. This categorization helps in understanding the content and context of the tweets or reviews, allowing for further analysis and interpretation. The categorization process provides a high-level overview of the prevalent subjects or discussions within the dataset, enabling researchers or analysts to focus on specific topics of interest or to explore patterns and trends within each category.

To categorize the reviews, two different processes are employed: fuzzy

Table 4.1. Top 20 features and their importance obtained from random forest classifier.

Feature	Importance
word_first	0.069157193
nb_words	0.032506769
doc2vec_vector_0	0.027300303
word_fever	0.026786426
doc2vec_vector_3	0.026066172
doc2vec_vector_4	0.02598418
doc2vec_vector_2	0.02547618
doc2vec_vector_1	0.024757602
word_pfizerbiontech	0.024465458
word_economical	0.021743698
word_new	0.018685477
word_vaccine	0.018394884
word_great	0.014174377
word_thanks	0.012901139
word_good	0.011893046
word_dose	0.011391914
word_get	0.009358373
word_pfizer	0.00811079
word_third	0.007627166
word_fast	0.00756942

string matching and angular similarity. These processes compare the similarity between each review and the index terms in different categories. The aim is to determine which category best matches the content of each review.

In the fuzzy string matching method, the similarity ratio is calculated by comparing the review with the index terms in each category. This method measures the degree of similarity between strings, taking into account differences in spelling, word order, and other factors. On the other hand, the angular similarity is determined using cosine similarity, which compares the angle between the word vector and the index term vectors in each category. This method calculates the similarity based on the orientation and magnitude of the vectors.

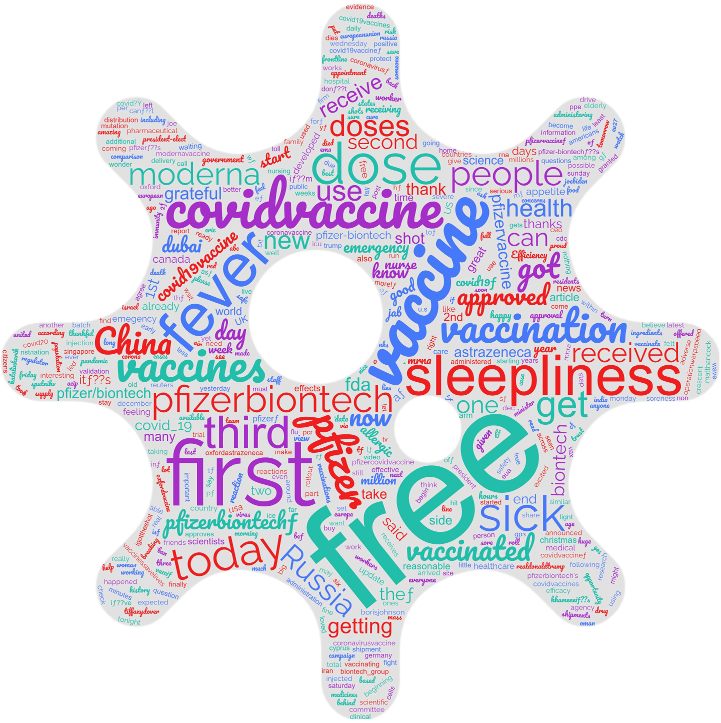


Figure 4.4. Wordcloud for determining the frequent words in vaccine review data

After obtaining similarity ratios from both fuzzy string matching and angular similarity, the average similarity value is computed for each review. This average represents the overall similarity between the review and the categories. The review is then assigned to the category with the highest average similarity score. By utilizing these processes, the reviews are effectively classified into categories based on their similarity to the index terms. This categorization enables a more organized and structured analysis of the reviews, facilitating further insights and interpretation.

An example of vaccine tweet/review with the average values of similarity with each of the category is provided in Figure 4.5.

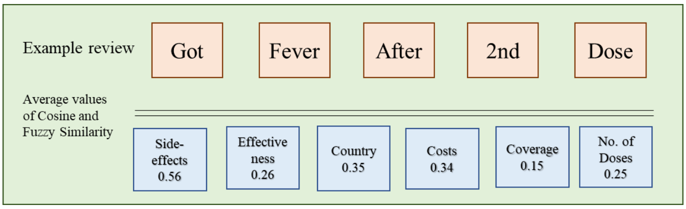


Figure 4.5. A review with the similarity values with each of the category

Based on Figure 4.5, it is evident that the reviews exhibit the highest similarity with the category labeled as "side effects". This suggests that the content and sentiment expressed in the reviews align closely with the topics and themes associated with side effects of a particular product, service, or experience. The maximum similarity with the "side effects" category indicates that a significant number of reviews within the dataset discuss or mention side effects in their content. This finding can be valuable for further analysis and decision-making, as it provides insights into the specific aspect or concern that users commonly associate with the product or service being reviewed.

The overall model Architecture for the review categorization process is shown in Figure 4.6.

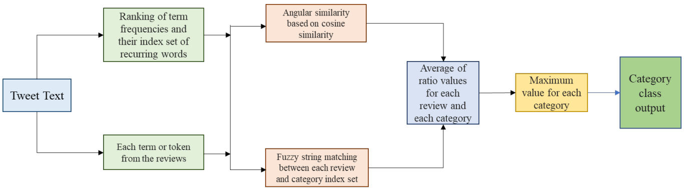

Tweet Text
Ranking of term frequencies and their index set of recurring words
Each term or token from the reviews
Angular similarity based on cosine similarity
Fuzzy string matching between each review and category index set
Average of ratio values for each review and each category
Maximum value for each category
Category class output

Figure 4.6. Architecture for the review categorization process

4.3 Results and Discussion

We used the sigma high performance architecture (explained in 3.3 and 3.4), for this work.

4.3.1 Comparison With the State-of-the-Art and Evaluation Matrices

We used following 4.2 hyperparameters for fine tuning:

Table 4.2. Hyperparameters for Sentiment Classification

Hyperparameter	**Values**
Model Name	CT-BERT
Max Sequence Length	128
Batch Size	32
Learning Rate	2e-6
Optimizer	AdamW
Number of Epochs	100
Warmup Proportion	0.1
Dropout Probability	0.1
Weight Decay	2e-4
Loss Function	Cross-Entropy

We have compared our sentiment analysis model with the other state-of-the-art models using the same dataset. The state-of-the-art models include BERT, [76], Random forest [23], Gated Recurrent Units (GRU) [66] and Long Short Term Memory (LSTM) [46]. To compare the performance of these models with our proposed model, we used the evaluation matrices of Accuracy, Precision, Recall and F1 measure. For the model evaluation, the dataset is partitioned into train, validation and test.

4.3.2 RQ1: How does the proposed ensemble approach, combining CT-BERT-CONVLayer_Fusion and Random Forest, contribute to accurate sentiment classification of COVID-19-related tweets?

A plot between training accuracy and validation accuracy per epochs in the CT-BERT_CONVLayerFusion model is shown in Figure 4.7

The observed plot clearly demonstrates a noticeable upward trajectory in both the training and validation accuracies during the final epochs. This indicates that the CT-BERT_CONVLayerFusion model does not suffer from overfitting, as there is no discernible decline in the test accuracy. The incorporation of Dropout layers in the BERT model plays a significant role in mitigating overfitting concerns. These Dropout layers selectively eliminate inputs during the training process, employing a probabilistic approach. By doing so, the model avoids excessive reliance on specific inputs within a given layer, thus effectively circumventing overfitting.

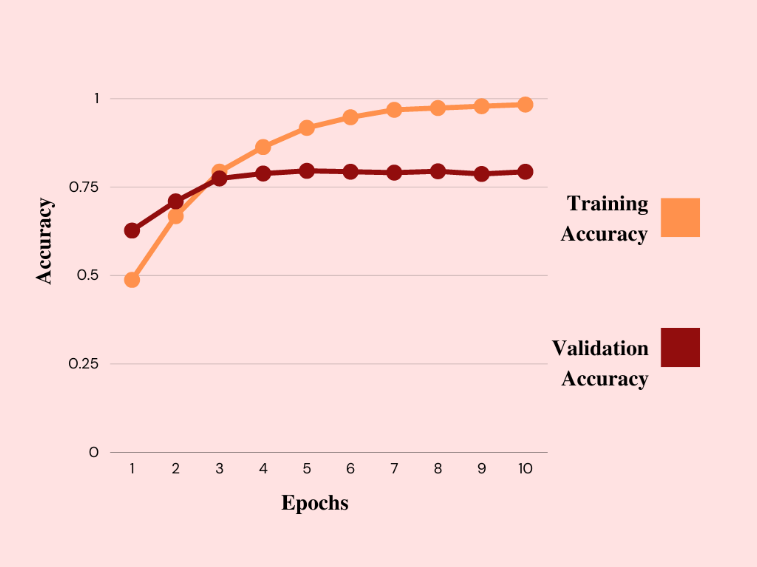


Figure 4.7. Graph of Training and Validation accuracy of our proposed the CT-BERT_CONVLayerFusion model

The comparison between our proposed Sentiment Analysis model and the state-of-the-art models for various sentiment categories, including strongly negative, mild negative, weakly negative, neutral, weakly positive, mild positive, and strongly positive, is illustrated in Figures 4.8, 4.9, 4.10, 4.11, 4.12, 4.13, and 4.14, respectively. Remarkably, our proposed model exhibits superior performance when compared to all other state-of-the-art models, achieving higher accuracy, precision, recall, and F1-measure scores. This exceptional outcome can be attributed to the fine-tuning step that we incorporated into our approach, which allows us to surpass conventional methods and achieve enhanced results.

By incorporating Convolutional layers and max pooling on the last four layers of the BERT model, we are able to capture information at multiple levels of granularity. The lower layers primarily capture local information, while the higher layers focus on global information. Through pooling across all four layers, we effectively capture information at both the local and global levels. This approach contributes to the improved results obtained in sentiment analysis tasks.

Moreover, it is worth noting that BERT is a large model with an extensive number of parameters, resulting in computationally expensive inference. However, by applying Convolutional layers and max pooling on the last four layers of the model, we can effectively reduce the dimensionality of the feature space. As a result, the inference process is accelerated without compromising the model's performance. This reduction in dimensionality provides a valuable trade-off between computational efficiency and model effectiveness.

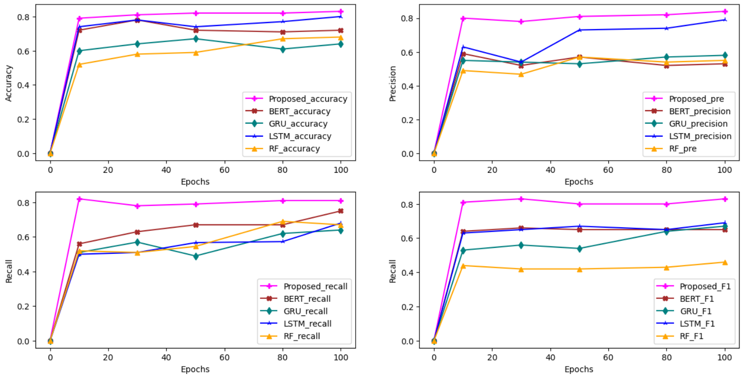


Figure 4.8. Sub-graphs of Accuracy, Precision, Recall and F1 showing the comparison of our proposed Sentiment Analysis model with state-of-the-art for strongly negative tweet classification.

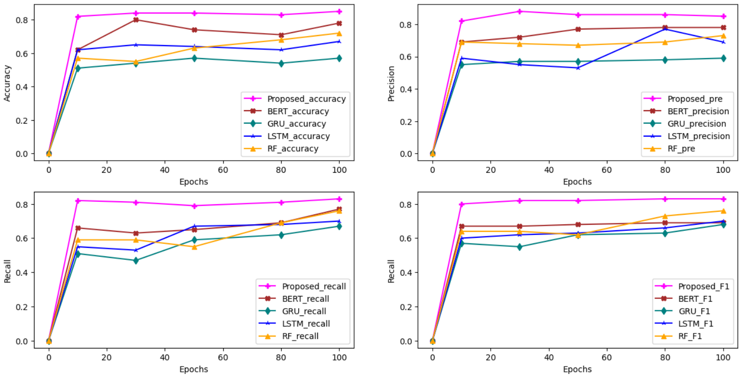


Figure 4.9. Sub-graphs of Accuracy, Precision, Recall and F1 showing the comparison of our proposed Sentiment Analysis model with state-of-the-art for mild negative tweet classification.

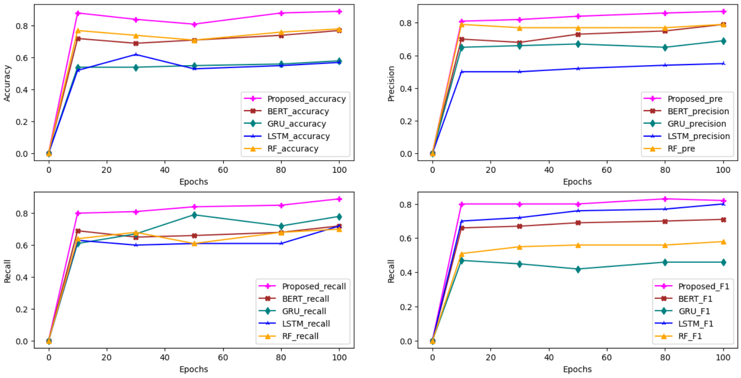


Figure 4.10. Sub-graphs of Accuracy, Precision, Recall and F1 showing the comparison of our proposed Sentiment Analysis model with state-of-the-art for weakly negative tweet classification.

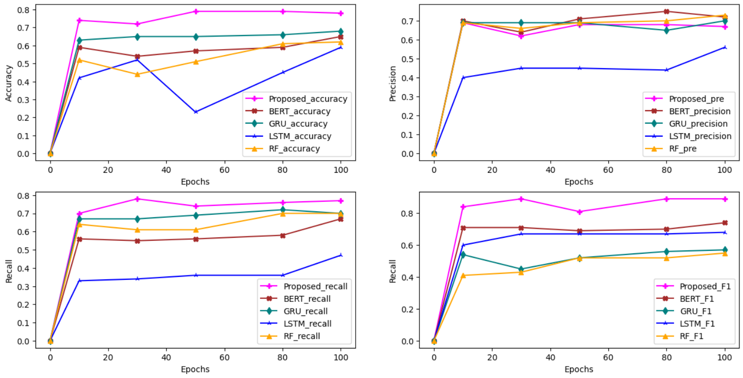


Figure 4.11. Sub-graphs of Accuracy, Precision, Recall and F1 showing the comparison of our proposed Sentiment Analysis model with state-of-the-art for neutral tweet classification.

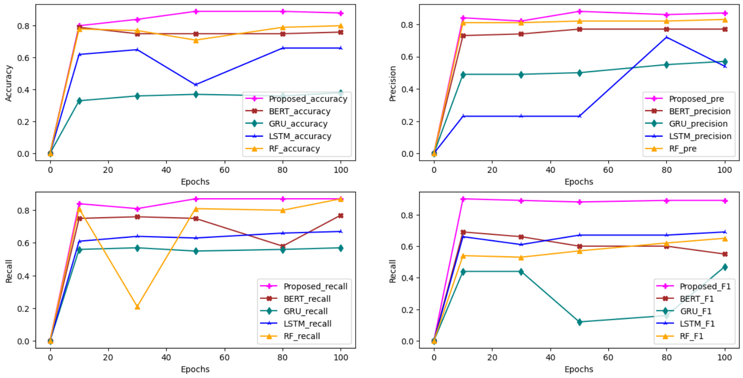


Figure 4.12. Sub-graphs of Accuracy, Precision, Recall and F1 showing the comparison of our proposed Sentiment Analysis model with state-of-the-art for weakly positive tweet classification.

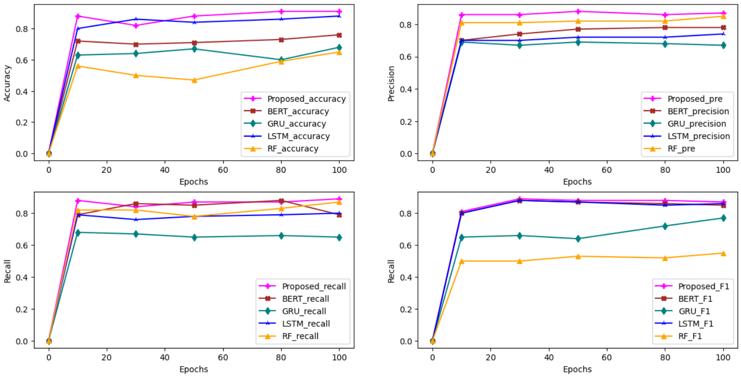


Figure 4.13. Sub-graphs of Accuracy, Precision, Recall and F1 showing the comparison of our proposed Sentiment Analysis model with state-of-the-art for mild positive tweet classification.

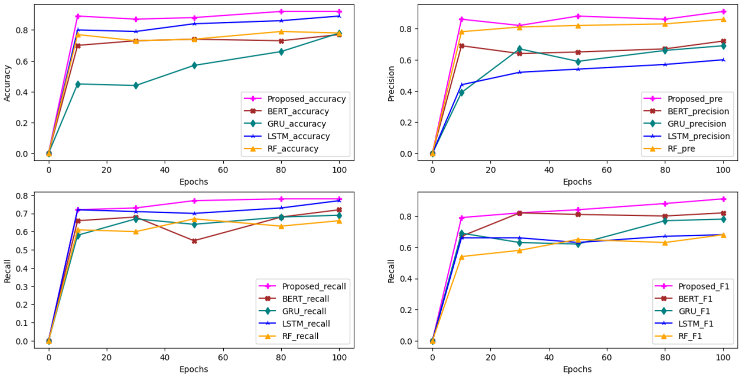


Figure 4.14. Sub-graphs of Accuracy, Precision, Recall and F1 showing the comparison of our proposed Sentiment Analysis model with state-of-the-art for strongly positive tweet classification.

The outstanding performance of our proposed BERT-based model can be attributed to several factors. Firstly, BERT benefits from pre-training on an extensive corpus of text data, enabling it to grasp the contextual nuances of language more effectively compared to models like LSTM and GRU, which are solely trained on specific tasks [18]. This pre-training allows BERT to develop a comprehensive understanding of language usage, leading to enhanced performance in various tasks.

Additionally, BERT's bi-directional nature plays a pivotal role in its success. By considering the entire sentence context in both forward and backward directions, BERT can capture a holistic understanding of the input text. This bi-directionality enables BERT to grasp the dependencies between words and capture intricate relationships within the sentence. Furthermore, BERT employs an attention mechanism that facilitates focused attention on the most pertinent words in the sentence while disregarding irrelevant ones. This attention mechanism plays a crucial role in enabling BERT to identify and capture the most significant features within the input text, ultimately enhancing its performance [43]. In summary, the combination of BERT's pre-training on extensive text data, its bi-directional nature, and its attention mechanism collectively contribute to the exceptional performance of our proposed BERT-based model in sentiment analysis and other related tasks.

The relatively lower performance of LSTM can be attributed to its unidirectional nature. LSTM processes sentences sequentially, either from left to right or right to left [74]. This unidirectional approach limits LSTM's ability to capture the intricate dependencies and contextual relationships between words in a sentence, as it does not have a comprehensive understanding of the entire sentence context. On the other hand, random forest models have demonstrated good performance in sentiment analysis tasks. However, their effectiveness may be limited when it comes to capturing the complex relationships between words and the context in which they are used. Random forest models typically operate by independently considering the features of each word without explicitly capturing the sequential and contextual information present in the sentence.

In contrast, BERT, as previously discussed, excels in capturing contextual information due to its pre-training on extensive text data, its bi-directional nature, and its attention mechanism. These factors allow BERT to effectively capture the intricate relationships between words and their context, leading to improved performance in sentiment analysis and similar tasks. In summary, the limitations of LSTM as a unidirectional model and the potential shortcomings of random forest models in capturing complex word relationships and contextual information contribute to their relatively lower performance compared to BERT-based models in sentiment analysis.

4.3.3 RQ2: To what extent does the sentiment analysis contribute to the categorization of COVID-19 vaccine-related tweets into specific predefined categories, and how does this enhance the recommendation system's comprehensiveness?

The various categories or tweets/reviews are given here:

- Side-Effects= "fever", "Sleepiness", "Loss of appetite", "body pain", "headache", "chills","fatigue","swelling","Swollen lymph nodes","muscle and joint pain"
- Effectiveness = "90 %", "80 %", "70 %", "60 %", "50 %", "40 %", "30 %", "20 %", "10 %"
- Country= "US", "UK","China","Germany","Belgium","Russia","India"
- Cost= "price", "amount", "rate", "cheap", "worth", 'money', "economical", "reasonable", "fee", "expensive", "charge","value"
- Coverage= "high", "medium", "low"
- Number of Doses= "one","two","three", "four","five"

The distribution of tweets based on the category type is reported in Figure 4.15.

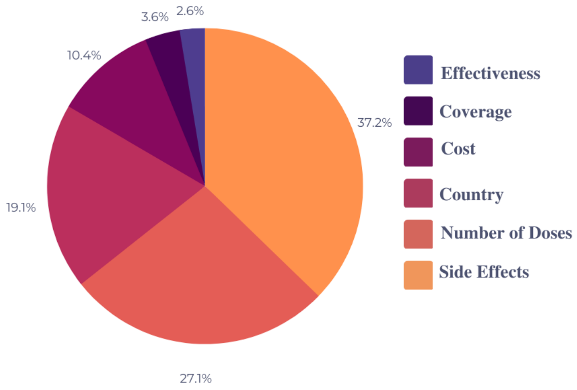


Figure 4.15. Pie Chart showing the distribution of tweets based on the category type

From the shown diagram, we can observe that the majority of the reviews are from the "Side Effects" and "Number of Doses" category. The latter information is important as it shows that the "Producer" of the vaccine that at the beginning of the vaccine use generated a lot of discussion quickly lost of relevance.

4.4 Main Contributions

Incorporating convolutional layers into BERT can offer several potential advantages:

- Local Context Capture: Convolutional layers excel at capturing local dependencies and patterns in data. By adding convolutional layers to BERT, you can potentially enhance its ability to capture local context within the input text. This can be especially useful for tasks where local information plays a significant role, such as named entity recognition or part-of-speech tagging.

- Parameter Efficiency: Convolutional layers are known for their parameter efficiency compared to fully connected layers. By incorporating convolutional layers into BERT, you can potentially reduce the number of parameters required in the model while still maintaining or even improving performance. This can be beneficial in scenarios where computational resources or memory limitations are a concern.

- Feature Extraction: Convolutional layers are effective at extracting low-level and mid-level features from the input data. By integrating convolutional layers into the last layers of BERT, you can potentially enable the model to learn more expressive and task-specific representations. This can enhance the model's ability to capture important features relevant to the downstream task.

- Generalization: The addition of convolutional layers can enhance the generalization ability of BERT. Convolutional layers are known to be robust to variations in input data, such as translations or slight shifts. By incorporating convolutional layers into BERT, the model may become more resilient to certain types of input perturbations, leading to improved generalization performance.

It's worth noting that the advantages of adding convolutional layers to BERT can vary depending on the specific task and dataset. Therefore, it is important to carefully evaluate the impact of such modifications through thorough experimentation and analysis.

Chapter 5

Geo-Spatial Analysis of COVID Vaccine Tweets

As previously defined, we set COVID-19 case study as an example of complex data. In this chapter[1], we applied modern GIS technologies to complex tweets dataset and visualize the current state of the COVID-19 disease and its behaviour on a large scale. Mapping can be used as a tool to investigate the relationship of disease with respect to its environment [91].

5.1 Methods

Geo Information Systems have many useful functions and tools which enable researchers to investigate the spatial aspects of the disease and help its monitoring and recovery. To analyze the disease spread dynamically, GIS has certain functions such as network analysis, buffer analysis, and statistical analysis which can estimate the future trend of disease with respect to its spatial aspects [10]. In this section, we describe our approach to refining the vaccine center indications taking into account the findings on the people's sentiment thus acting as a kind of specialized recommender system, as this kind of systems are extensively studied nowadays especially when they take advantage of well established tools like GIS.

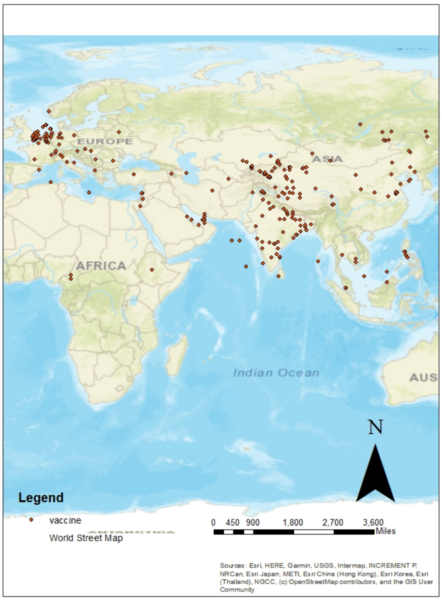
[1] The content of this Chapter is mainly based on References [76] and [80] of the reference list.

5.1.1 Vaccine hesitancy due to access issues.

Based on the results of the word clouds presented in Section 3.4.3 in Chapter 3, it can be noted that sentiment analysis demonstrates a low willingness to vaccinate. Therefore, the use of geospatial analysis to identify potential barriers to vaccination can be fruitful. For some individuals, vaccine hesitancy may be driven by barriers to accessing vaccines, such as lack of transportation, long wait times, or difficulty scheduling appointments. The report by the WHO Strategic Advisory Group of Experts (SAGE) on Immunization identifies confidence (lack of trust towards vaccine providers), complacency (unable to understand the importance of a vaccine for a particular disease), and convenience (access to vaccines i.e. physical availability, geographical accessibility, and the ability to understand because of issues with language or health literacy affecting uptake) as the three main factors influencing vaccine hesitancy [8]. Geospatial approaches can address the third factor i.e., convenience or access to vaccination by suggesting proper vaccination centers to people based on sentiment analysis results in a given area.

5.1.2 Geo-Coding and Visualization of Data:

We partitioned our dataset into subsets and applied geocoding to each instance. Geocoding is the process of converting the address of a location into its respective geographic coordinates. We used the geoPy library in Python for geocoding, which utilizes third-party geocoders to locate geographic coordinates [15]. After this step, we visualized the geocoded vaccination data on a map surface using ArcGIS 10.5. In Figure 5.1, vaccine distribution around the world is shown.



Sources: Esri, HERE, Garmin, USGS, Intermap, INCREMENT P, NRCan, Esri Japan, METI, Esri China (Hong Kong), Esri Korea, Esri (Thailand), NGCC, (c) OpenStreetMap contributors, and the GIS User Community

Figure 5.1. Vaccine data Visualization

5.1.3 Geographical correlation:

Discovering relationships among features in a dataset is crucial for identifying spatial patterns. In more detail, feature points can be spatially clustered, random, or dispersed. The null hypothesis assumes that features are entirely randomly distributed. The pattern analysis we performed returned a p-value and z-score, which are used to reject the null hypothesis. Therefore, if the null hypothesis is falsified, there exists a relationship among the features. This relationship may indicate clustering or dispersion. If a clustered relationship exists, it shows high geographical associativity among the features.

P-value and Z-score Computation.

P-value is a measure of the probability of obtaining a result as extreme as, or more extreme than, the observed result. It shows the spatial pattern of the random process and ranges between 0 and 1. Smaller values indicate that the pattern is not random. Z-score measures the standard deviation, which computes how the data are distributed with respect to the mean value. For example, a Z-score value of 0 shows that the observed value is equal to the average value, while 2.5 means that it is 2.5 away from the average. Z-scores can be positive or negative. A high (positive) z-score indicates that the data point is far above the mean, while a low (negative) z-score means that the data point is far below the mean. A confidence range is assigned to the range of P-values and Z-scores as shown in Table 5.1 [54]

Table 5.1. Confidence level associated with P-values and Z-scores

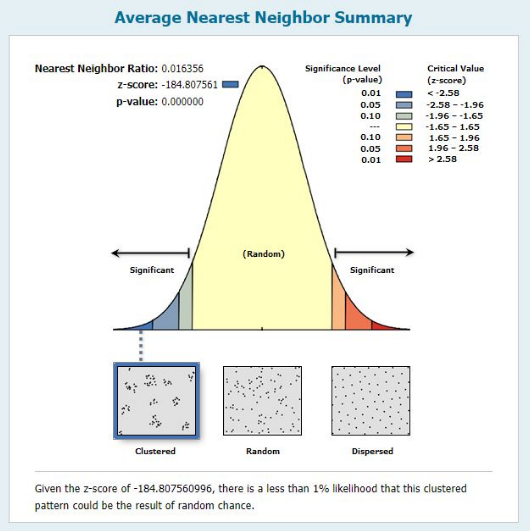
P-value	Z-Score	Confidence Level
<0.10	<-1.65 or > +1.65	90 %
<0.05	<-1.96 or > +1.96	95 %
<0.01	<-2.58 or > +2.58	99 %

The P-values and Z-scores are computed together to test the null hypothesis. If the Z-score is very high or very low combined with a small P-value, the null hypothesis is rejected [54].

Average Nearest Neighbor:

We leveraged the ANN (Average Nearest Neighbor) tool of ArcGIS to identify the spatial relationship among the points in the vaccine dataset. The ANN is defined as the ratio of the observed average distance (DO) to the expected average distance (DE), and can be computed using Equation 5.1.

$$ANN = DO/DE \tag{5.1}$$



Average Nearest Neighbor Summary

Nearest Neighbor Ratio: 0.016356
z-score: -184.807561
p-value: 0.000000

Significance Level (p-value)		Critical Value (z-score)
0.01	🟦	< -2.58
0.05	🟦	-2.58 - -1.96
0.10	▨	-1.96 - -1.65
---	🟨	-1.65 - 1.65
0.10	🟧	1.65 - 1.96
0.05	🟧	1.96 - 2.58
0.01	🟥	> 2.58

Given the z-score of -184.807560996, there is a less than 1% likelihood that this clustered pattern could be the result of random chance.

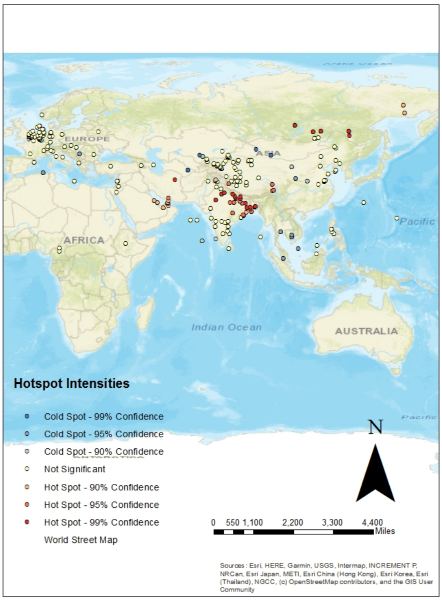
Figure 5.2. Spatial correlation among the data features

If the value of ANN is less than 1, it indicates that the dataset is clustered [54]. In our case, we obtained a value that is less than 1, confirming that our data are clustered, as shown in Figure 5.2.

Furthermore, we can observe in Figure 5.2 that the Z-score obtained for our dataset is -184.807561, while the P-value is 0.000000. Upon examining the values of the z-score and p-values in Table 5.1, we can see that the P-value puts us in a 99% confidence level, while the Z-score puts us in a 95% confidence level. Therefore, we can conclude that our dataset is clustered.

5.1.4 Hotspot Analysis

Hotspot shows the geographical areas [75] where the vaccine sentiment polarity is high in rate while cold spots shows the areas with less vaccine sentiment polarity. Creating hotspots in maps help to better investigate the sentiments of the people toward vaccine. We used the Hotspot analysis tool of ArcGIS software for this purpose. It works based on the polarity values of each tweet.


EUROPE
ASIA
AFRICA
Pacific
Indian Ocean
AUSTRALIA
Pacific
Hotspot Intensities
Cold Spot - 99% Confidence
Cold Spot - 95% Confidence
Cold Spot - 90% Confidence
Not Significant
Hot Spot - 90% Confidence
Hot Spot - 95% Confidence
Hot Spot - 99% Confidence
World Street Map
N
0 550 1,100 2,200 3,300 4,400
Miles

Sources: Esri, HERE, Garmin, USGS, Intermap, INCREMENT P, NRCan, Esri Japan, METI, Esri China (Hong Kong), Esri Korea, Esri (Thailand), NGCC, (c) OpenStreetMap contributors, and the GIS User Community

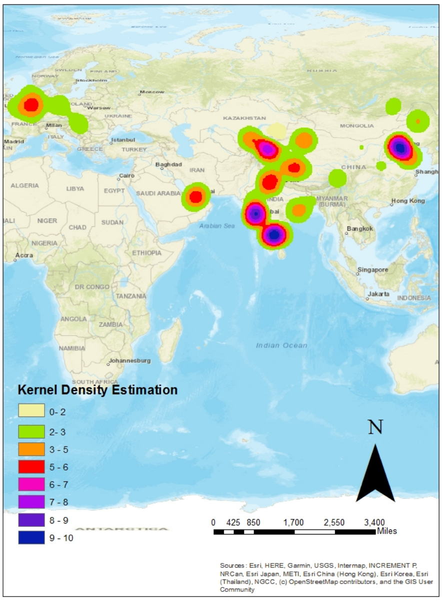
Figure 5.3. Hotspot analysis of COVID-Vaccine tweets

It is clearly visible from the Figure 5.3 that Asia countries such as India, Saudi Arabia are showing more positive attitude towards the vaccines while Europe is behaving neutral during vaccine. China and few other countries are showing negative sentiments for vaccine. According to a systematic review of vaccine acceptance rates in [63], higher-income, age and gender are the main reason behind the different behaviors of people in different regions of the world.

5.1.5 Analysis Using Kernel Density Estimation:

Point density can be envisioned as the series of circles around each feature point, and the density being calculated as the number of circles being over-lapped [17]. Kernel density can be envisioned as putting the blob of ice-cream on the top of each feature point and then the density function be the measuring the height of the accumulated blobs. It interprets

COVID-19 vaccine data and extract valuable information for COVID-19 modelling. We used the kernel density tool of ArcGIS to perform density analysis. Figure 5.4 represents the clusters formed using kernel density over COVID-19 Vaccine dataset.



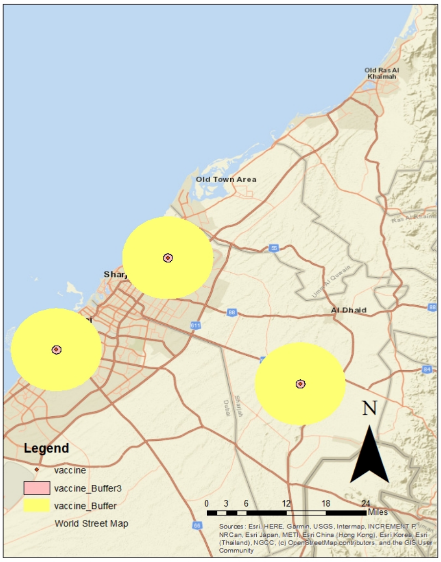
Sources: Esri, HERE, Garmin, USGS, Intermap, INCREMENT P, NRCan, Esri Japan, METI, Esri China (Hong Kong), Esri Korea, Esri (Thailand), NGCC, (c) OpenStreetMap contributors, and the GIS User Community

Figure 5.4. Kernel Density Estimation of COVID-Vaccine tweets

Figure 5.4 gives more closer overview of the people's sentiment all over the globe. It also shows that the more positive sentiment polarity is found in India and Europe behave neutrally in this context.

5.1.6 Buffering

We discussed in previous sections that a possible obstacle to getting vaccinated may be mobility, as defined in Section 5.1.1. Geo-spatial techniques, like buffering, can be efficiently used to address allocation problems [35, 71]. We performed buffering using ArcGIS to find the suitability of vaccination centers with respect to people's locations. The buffering analysis of our data is reported in Figure 5.5, which shows the willingness of people to move for vaccination. We assumed that vaccination centers must be within 10 km of the user's address. Indeed, several scenarios were run with different distance values, and 10 km resulted in the most favored distance to encourage people to get vaccinated



Sources: Esri, HERE, Garmin, USGS, Intermap, INCREMENT P, NRCan, Esri Japan, METI, Esri China (Hong Kong), Esri Korea, Esri (Thailand), NGCC, (c) OpenStreetMap contributors, and the GIS User Community

Figure 5.5. Results of Buffering on Vaccination dataset

Figure 5.5 shows the results we obtained by running buffering over the vaccination dataset. We can observe two types of buffers in the diagram. The pink buffer visualizes the nearest possible vaccination centers, while the yellow buffer shows the results on an international scale.

Chapter 6

Conclusions

Social media-based sentiment analysis of people's feelings about vaccines is a useful and cost-effective way to design policies for vaccination campaigns. As a matter of fact, vaccines were developed to control the spread of COVID-19 worldwide, but vaccine hesitancy seems to be an even bigger challenge than COVID-19. In this thesis, we identify people's reactions during the vaccination phase.

1. We used X app (former Twitter) data and applied various pre-processing steps i.e. data cleaning, URLs removal, stopwords removal, punctuation removal,

2. We found out the polarity of the tweets using textBlob() function.

3. We categorized tweets into seven categories, considering tweet polarity values. These categories are weakly negative, mild negative, strongly negative, neutral, weak positive, mild positive and strongly positive.

4. We proposed a BERT+NBSVM classification model for the classification of positive and negative sentiments.

Traditional computing architectures could exhibit limited performances as regards processing power and efficiency when executing complex computing tasks required by modern applications such as Artificial Intelligence. As a result, high-performance architectures are used in this research,to provide faster and more efficient computing capabilities.

After the devastation and havoc of COVID-19, it is important to develop systems that can help health care practitioners in combating COVID-19 and to provide patient or each individual with personalized vaccine recommendations. Helping an individual to choose a proper vaccine based on his/her medical history and other requirements from the online feedback/reviews given by other people, contributes to the research field called vaccine recommendation system. This systems helps customers to intake proper and more informed vaccination based on their preferences and beliefs that they provide in the form of queries. For this purpose:

1. We designed a Covid-19 vaccine recommender system based on sentiment analysis.

2. We proposed a novel sentiment analysis model that classify the tweets into seven different sentiment categories thus refining the usual classification in two categories (positive and negative).

3. We grouped the tweets into different aspect based categories.

4. Based on the user query, an appropriate vaccine along with their reviews is selected.

Indeed, the use of advance deep learning methods to analyse public information can guide people for a more conscious choice regarding any medical device that represent our future challenge. Our work also focuses on the usage of geo-spatial approaches to identify the geo-spatial patterns in the vaccination data. We performed buffering to suggest proper vaccination centers based on sentiment analysis. Thus, policymakers can benefit from our methods by analyzing people's concerns and understanding their mindset to improve proper planning to inform people about vaccines, identify misinformation or rumors spreading across the country, and launch ad-hoc campaigns suited to avoid confusion on this important topic.

In the context of the extensive comparison conducted in Chapter 2, it is essential to acknowledge the diverse landscape of sentiment analysis during the COVID-19 pandemic, as illustrated in Table 2.1. While our focus has been on developing an innovative recommender system, it is imperative to recognize the broader challenges and future directions highlighted in the comparison. The limitations identified, such as concerns related to data size, language variations, and real-time classification, resonate with the complexities of sentiment analysis. Our recommender system, although a significant contribution, is not exempt from these challenges. However, our approach stands as a testament to the ongoing efforts to enhance sentiment analysis methodologies. By leveraging advanced techniques and incorporating diverse data sources, including those used in the comparison, we aim to mitigate these challenges and provide users with a robust and effective tool for informed decision-making. The lessons learned from the broader sentiment analysis landscape have been instrumental in shaping the development of our recommender system, contributing to its adaptability and potential impact in real-world scenarios.

6.1 Limitations and Future Works

While the proposed techniques in this work have shown promising results, there are certain limitations that warrant consideration. Firstly, the reliance on social media data, particularly from platforms like X app (formerly Twitter), introduces challenges related to data noise, unstructured content, and potential biases in user sentiments. The sentiment analysis, though effective, may be influenced by the dynamic nature of social media discussions and the evolving sentiment landscape. Additionally, the proposed ensemble model, while robust, may have computational demands that need optimization for scalability. The Geo-Spatial approaches utilized for geographical sentiment analysis may also face limitations in terms of data coverage and accuracy.

Future work could focus on addressing these limitations by exploring more advanced techniques for noise reduction, incorporating real-time sentiment dynamics, and enhancing the scalability of the proposed models. Furthermore, efforts to improve the interpretability and explainability of AI results in the context of complex data could contribute to a more transparent and trustworthy analysis. Research avenues also include extending the proposed sentiment analysis framework to accommodate evolving sentiment patterns and exploring novel data sources to enrich the analysis with diverse perspectives. These endeavors can collectively contribute to advancing the understanding and application of sentiment analysis in the domain of complex data.

Author's Publications

1. Umair, Areeba, and Elio Masciari. "Sentimental and spatial analysis of covid-19 vaccines tweets." Journal of Intelligent Information Systems 60.1 (2023): 1-21.

2. Umair, Areeba, Elio Masciari, and Muhammad Habib Ullah. "Vaccine sentiment analysis using BERT+ NBSVM and geo-spatial approaches." The Journal of Supercomputing (2023): 1-31.

3. Umair, Areeba, 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.

4. Umair, Areeba, et al. "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.

5. Umair, Areeba, 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.

6. Umair, Areeba, and Elio Masciari. "Human sentiments monitoring during COVID-19 using AI-based modeling." Procedia Computer Science 203 (2022): 753-758.

7. Umair, Areeba, and Elio Masciari "Sentiment Analysis using Improved CT-BERT_CONVLayer Fusion Model for COVID-19 Vaccine Recommendation" Journal, 2023 (to be submitted)

Bibliography

[1] Noora Abdulrahman and Wala Abedalkhader. KNN Classifier and Naive Bayse Classifier for Crime Prediction in San Francisco Context. *Int. J. Database Manag. Syst.*, 9(4):1–9, 2017.

[2] Hassan Adamu, Syaheerah Lebai Lutfi, Nurul Hashimah Ahamed Hassain Malim, Rohail Hassan, Assunta Di Vaio, and Ahmad Sufril Azlan Mohamed. Framing twitter public sentiment on Nigerian government COVID-19 palliatives distribution using machine learning. *Sustain.*, 13(6), 2021.

[3] Bahaulddin Nabhan Adday, Faris Ali Jasim Shaban, Mohammed Rasool Jawad, Refed Adnan Jaleel, and Musadaq Mahir Abdel Zahra. Enhanced vaccine recommender system to prevent covid-19 based on clustering and classification. In *2021 international conference on engineering and emerging technologies (iceet)*, pages 1–6. IEEE, 2021.

[4] Ashi Agarwal, Basant Agarwal, Priyanka Harjule, and Ajay Agarwal. *Mental Health Analysis of Students in Major Cities of India During COVID-19.* Springer Singapore, 2021.

[5] V. Ajantha Devi and Anand Nayyar. *Evaluation of Geotagging Twitter Data Using Sentiment Analysis During COVID-19*, volume 166. Springer Singapore, 2021.

[6] Tahani Almanie, Rsha Mirza, and Elizabeth Lor. Crime Prediction Based on Crime Types and Using Spatial and Temporal Criminal Hotspots. *Int. J. Data Min. Knowl. Manag. Process*, 5(4):1–19, 2015.

[7] Elham Asani, Hamed Vahdat-Nejad, and Javad Sadri. Restaurant recommender system based on sentiment analysis. *Machine Learning with Applications*, 6:100114, 12 2021.

[8] Helen Bedford, Katie Attwell, Margie Danchin, Helen Marshall, Paul Corben, and Julie Leask. Vaccine hesitancy, refusal and access barriers: The need for clarity in terminology. *Vaccine*, 36(44):6556–6558, 2018.

[9] Merlijn Blaauw and Jordi Bonada. Sequence-to-Sequence Singing Synthesis Using The Feed-Forward Transformer. *arXiv*, pages 7229–7233, 2019.

[10] Maged N Kamel Boulos and Estella M Geraghty. Geographical tracking and mapping of coronavirus disease COVID - 19 / severe acute respiratory syndrome coronavirus 2 (SARS - CoV - 2) epidemic and associated events around the world : how 21st century GIS technologies are supporting the global fight ag. *Int. J. Health Geogr.*, pages 1–12, 2020.

[11] Nunziato Cassavia and Elio Masciari. Sigma: a scalable high performance big data architecture. In *2021 29th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*, pages 236–239. IEEE, 2021.

[12] Tanmoy Chakraborty, Kai Shu, H Russell Bernard, Huan Liu, and Md Shad Akhtar. *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers*, volume 1402. Springer Nature, 2021.

[13] Yuanhang Chen, Gaoliang Peng, Zhiyu Zhu, and Sijue Li. A novel deep learning method based on attention mechanism for bearing remaining useful life prediction. *Appl. Soft Comput. J.*, 86:105919, 2020.

[14] Nalini Chintalapudi, Gopi Battineni, and Francesco Amenta. Sentimental Analysis of COVID-19 Tweets Using Deep Learning Models. *Infect. Dis. Rep.*, 13(2):329–339, 2021.

[15] Miyoung Chong and Haihua Chen. Racist framing through stigmatized naming: A topical and geo-locational analysis of# chinavirus and# chinesevirus on twitter. *Proceedings of the Association for Information Science and Technology*, 58(1):70–79, 2021.

[16] Wen Ying Sylvia Chou and Alexandra Budenz. Considering Emotion in COVID-19 Vaccine Communication: Addressing Vaccine Hesitancy and Fostering Vaccine Confidence. *Health Commun.*, 35(14):1718–1722, 2020.

[17] Peter D. Haan. On the use of density kernels for concentration estimations within particle and puff dispersion models. *Atmos. Environ.*, 33(13):2007–2021, 1999.

[18] Zhuyun Dai and Jamie Callan. Deeper text understanding for ir with contextual neural language modeling. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 985–988, 2019.

[19] Cach N. Dang, María N. Moreno-García, and Fernando De la Prieta. An approach to integrating sentiment analysis into recommender systems. *Sensors*, 21, 8 2021.

[20] Subasish Das and Anandi Dutta. Characterizing public emotions and sentiments in COVID-19 environment: A case study of India. *J. Hum. Behav. Soc. Environ.*, 31(1-4):1–14, 2020.

[21] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, 1:4171–4186, 2019.

[22] Linhao Dong, Shuang Xu, and Bo Xu. SPEECH-TRANSFORMER : A NO-RECURRENCE SEQUENCE-TO-SEQUENCE MODEL FOR SPEECH RECOGNITION Institute of Automation , Chinese Academy of Sciences , China University of Chinese Academy of Sciences , China. *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pages 5884–5888, 2018.

[23] M Ali Fauzi. Random forest approach fo sentiment analysis in indonesian. *Indones. J. Electr. Eng. Comput. Sci*, 12:46–50, 2018.

[24] S. W. Flint, A. Piotrkowicz, and K. Watts. Use of Artificial Intelligence to understand adults' thoughts and behaviours relating to COVID-19. *Perspect. Public Health*, XX(X):1–8, 2021.

[25] F. A. Furfari(tony). The Transformer. *IEEE Ind. Appl. Mag.*, 8(1):8–15, 2002.

[26] Joma George, Shintu Mariam Skariah, and T Aleena Xavier. Role of contextual features in fake news detection: a review. In *2020 international conference on innovative trends in information technology (ICITIIT)*, pages 1–6. IEEE, 2020.

[27] Kamaran H. Manguri, Rebaz N. Ramadhan, and Pshko R. Mohammed Amin. Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks. *Kurdistan J. Appl. Res.*, pages 54–65, 2020.

[28] Hao Huang, Zongchao Peng, Hongtao Wu, and Qihui Xie. A big data analysis on the five dimensions of emergency management information in the early stage of COVID-19 in China. *J. Chinese Gov.*, 5(2):213–233, 2020.

[29] Man Hung, Evelyn Lauren, Eric S. Hon, Wendy C. Birmingham, Julie Xu, Sharon Su, Shirley D. Hon, Jungweon Park, Peter Dang, and Martin S. Lipsky. Social network analysis of COVID-19 sentiments: Application of artificial intelligence. *J. Med. Internet Res.*, 22(8):1–13, 2020.

[30] Michele Ianni, Elio Masciari, Giuseppe M. Mazzeo, Mario Mezzanzanica, and Carlo Zaniolo. Fast and effective big data exploration by clustering. *Future Gener. Comput. Syst.*, 102:84–94, 2020.

[31] Irem Islek and Sule Gunduz Oguducu. A hierarchical recommendation system for e-commerce using online user reviews. *Electronic Commerce Research and Applications*, 52:101131, 2022.

[32] Hamed Jelodar, Yongli Wang, Rita Orji, and Hucheng Huang. Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach. *arXiv*, 24(10):2733–2742, 2020.

[33] Akbar Karimi, Leonardo Rossi, and Andrea Prati. Improving bert performance for aspect-based sentiment analysis. *arXiv preprint arXiv:2010.11731*, 10 2020.

[34] L Kavisankar, S Balasubramani, D John Arvindhar, and Rajkumar Krishan. Scenario based vaccine status monitoring and recommendation system for covid-19 vaccination. *Journal of Management Information and Decision Sciences*, 24:1–7, 2021.

[35] Shahid Nawaz Khan, Kamran Mir, Ali Tahir, Arshad Awan, Zaib Un Nisa, and Syeda Areeba Gillani. Allocation of tutors and study centers in distance learning using geospatial technologies. *ISPRS International Journal of Geo-Information*, 7(5):185, 2018.

[36] Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics*, 11(1):141, 2022.

[37] Jay Kreps. Questioning the lambda architecture. *Online article, July*, 205:18–34, 2014.

[38] Jiamin Liu, Tao Wang, Jiting Li, Jingbo Huang, Feng Yao, and Renjie He. A data-driven analysis of employee promotion: The role of the position of organization. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 4056–4062. IEEE, 2019.

[39] Yang Liu and Mirella Lapata. Hierarchical transformers for multi-document summarization. *ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf.*, pages 5070–5081, 2020.

[40] Jie Lu, Dianshuang Wu, Mingsong Mao, Wei Wang, and Guangquan Zhang. Recommender system application developments: a survey. *Decision support systems*, 74:12–32, 2015.

[41] Yi Luo and Xiaowei Xu. Comparative study of deep learning models for analyzing online restaurant reviews in the era of the COVID-19 pandemic. *Int. J. Hosp. Manag.*, 94(December 2020):102849, 2021.

[42] May Oo Lwin, Jiahui Lu, Anita Sheldenkar, Peter Johannes Schulz, Wonsun Shin, Raj Gupta, and Yinping Yang. Global sentiments surrounding the COVID-19 pandemic on Twitter: Analysis of Twitter trends. *JMIR Public Heal. Surveill.*, 6(2):1–4, 2020.

[43] Fanqi Meng, Shuaisong Yang, Jingdong Wang, Lei Xia, and Han Liu. Creating knowledge graph of electric power equipment faults based on bert–bilstm–crf model. *Journal of Electrical Engineering & Technology*, 17(4):2507–2516, 2022.

[44] Md Mokhlesur Rahman, G. G.Md Nawaz Ali, Xue Jun Li, Kamal Chandra Paul, and Peter H.J. Chong. Twitter and Census Data Analytics to Explore Socioeconomic Factors for Post-COVID-19 Reopening Sentiment. *arXiv*, 2020.

[45] Niko Moritz, Takaaki Hori, and Jonathan Le Roux. Streaming Automatic Speech Recognition With the Transformer Model. *arXiv*, pages 6074–6078, 2020.

[46] GSN Murthy, Shanmukha Rao Allu, Bhargavi Andhavarapu, Mounika Bagadi, and Mounika Belusonti. Text based sentiment analysis using lstm. *Int. J. Eng. Res. Tech. Res*, 9(05), 2020.

[47] Martin Müller, Marcel Salathé, and Per E Kummervold. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *preprint arXiv:2005.07503*, 5 2020.

[48] Abdou Nagy and Bader Alhatlani. An overview of current covid-19 vaccine platforms. *Computational and structural biotechnology journal*, 19:2508–2517, 2021.

[49] Duduzile Ndwandwe and Charles S Wiysonge. Covid-19 vaccines. *Current opinion in immunology*, 71:111–116, 2021.

[50] László Nemes and Attila Kiss. Social media sentiment analysis based on COVID-19. *J. Inf. Telecommun.*, 5(1):1–15, 2021.

[51] Thu T. Nguyen, Shaniece Criss, Pallavi Dwivedi, Dina Huang, Jessica Keralis, Erica Hsu, Lynn Phan, Leah H. Nguyen, Isha Yardi, M. Maria Glymour, Amani M. Allen, David H. Chae, Gilbert C. Gee, and Quynh C. Nguyen. Exploring U.S. shifts in anti-Asian sentiment with the emergence of COVID-19. *Int. J. Environ. Res. Public Health*, 17(19):1–13, 2020.

[52] Niko Pajkovic. Algorithms and taste-making: Exposing the netflix recommender system's operational logics. *Convergence*, 28(1):214–235, 2022.

[53] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Improving Language Understanding by. *OpenAI*, pages 1–10, 2018.

[54] Eric Pimpler. *Spatial analytics with ArcGIS*. Packt Publishing Ltd, Birmingham, England, 2017.

[55] Marco Pota, Mirko Ventura, Rosario Catelli, and Massimo Esposito. An effective bert-based pipeline for twitter sentiment analysis: A case study in Italian. *Sensors (Switzerland)*, 21(1):1–21, 2021.

[56] S. V. Praveen, Rajesh Ittamalla, and Gerard Deepak. Analyzing Indian general public's perspective on anxiety, stress and trauma during Covid-19 - A machine learning study of 840,000 tweets. *Diabetes Metab. Syndr. Clin. Res. Rev.*, 15(3):667–671, 2021.

[57] SV Praveen, Rajesh Ittamalla, and Gerard Deepak. Analyzing the attitude of indian citizens towards covid-19 vaccine–a text analytics study. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 15(2):595–599, 2021.

[58] Eryka Probierz, Adam Gałuszka, and Tomasz Dzida. Twitter text data from #Covid-19: Analysis of changes in time using exploratory sentiment analysis. *J. Phys. Conf. Ser.*, 1828(1), 2021.

[59] Philipp Probst, Marvin N Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, 9(3):e1301, 2019.

[60] Supriya Raheja and Anjani Asthana. Sentimental analysis of twitter comments on COVID-19. *Proc. Conflu. 2021 11th Int. Conf. Cloud Comput. Data Sci. Eng.*, pages 704–708, 2021.

[61] Rucha Hemant Rangnekar, Khyati Pradeep Suratwala, Sanjana Krishna, and Sudhir Dhage. Career Prediction Model Using Data Mining And Linear Classification. In *Fourth Int. Conf. Comput. Commun. Control Autom.*, pages 1–6, 2018.

[62] Furqan Rustam, Madiha Khalid, Waqar Aslam, Vaibhav Rupapara, Arif Mehmood, and Gyu Sang Choi. A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *PLoS One*, 16(2):1–23, 2021.

[63] Malik Sallam. Covid-19 vaccine hesitancy worldwide: a concise systematic review of vaccine acceptance rates. *Vaccines*, 9(2):160, 2021.

[64] Jim Samuel, G. G.Md Nawaz Ali, Md Mokhlesur Rahman, Ek Esawi, and Yana Samuel. COVID-19 public sentiment insights and machine learning for tweets classification. *Inf.*, 11(6):1–22, 2020.

[65] Jim Samuel, Md Mokhlesur Rahman, G. G.Md Nawaz Ali, Yana Samuel, Alexander Pelaez, Peter Han Joo Chong, and Michael Yakubov. Feeling Positive about Reopening? New Normal Scenarios from COVID-19 US Reopen Sentiment Analytics. *IEEE Access*, 8:142173–142190, 2020.

[66] Yunus Santur. Sentiment analysis based on gated recurrent unit. In *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, pages 1–5. IEEE, 2019.

[67] Md Shahriare Satu, Md Imran Khan, Mufti Mahmud, Shahadat Uddin, Matthew A. Summers, Julian M.W. Quinn, and Mohammad Ali Moni. TClustVID: A novel machine learning classification model to investigate topics and sentiment in COVID-19 tweets. *medRxiv*, 2020.

[68] Patrick Schratz, Jannes Muenchow, Eugenia Iturritxa, Jakob Richter, and Alexander Brenning. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, 406:109–120, 2019.

[69] Holly Seale, Anita E. Heywood, Julie Leask, Meru Sheel, David N. Durrheim, Katarzyna Bolsewicz, and Rajneesh Kaur. Examining Australian public perceptions and behaviors towards a future COVID-19 vaccine. *medRxiv*, pages 1–9, 2020.

[70] Adnan Muhammad Shah, Xiangbin Yan, Abdul Qayyum, Rizwan Ali Naqvi, and Syed Jamal Shah. Mining topic and sentiment dynamics in physician rating websites during the early wave of the COVID-19 pandemic: Machine learning approach. *Int. J. Med. Inform.*, 149(February), 2021.

[71] Shahrooz Shahparvari, Masih Fadaki, and Prem Chhetri. Spatial accessibility of fire stations for enhancing operational response in melbourne. *Fire safety journal*, 117:103149, 2020.

[72] Bonggun Shin, Sungsoo Park, Keunsoo Kang, and Joyce C. Ho. Self-attention based molecule representation for predicting drug-target interaction. *arXiv*, pages 1–18, 2019.

[73] Mrityunjay Singh, Amit Kumar Jakhar, and Shivam Pandey. Sentiment analysis on the impact of coronavirus in social life using the BERT model. *Soc. Netw. Anal. Min.*, 11(1):1–11, 2021.

[74] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1441–1450, 2019.

[75] Donna R Tabangin, Jacqueline C Flores, and Nelson F Emperador. Investigating Crime Hotspot Places and their Implication to Urban Environmental Design : A Geographic Visualization and Data Mining Approach. *Int. J. Hum. Soc. Sci.*, 2(12):4004–4012, 2008.

[76] Areeba Umair and Elio Masciari. Sentimental and spatial analysis of covid-19 vaccines tweets. *Journal of Intelligent Information Systems*, pages 1–21, 2022.

[77] Areeba Umair and Elio Masciari. Using high performance approaches to covid-19 vaccines sentiment analysis. In *2022 30th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*, pages 197–204. IEEE, 2022.

[78] Areeba Umair, Elio Masciari, and Muhammad Habib Habib Ullah. Sentimental analysis applications and approaches during covid-19: A survey. In *25th International Database Engineering & Applications Symposium*, IDEAS 2021, page 304–308, New York, NY, USA, 2021. Association for Computing Machinery.

[79] Areeba Umair, Elio Masciari, Giusi Madeo, and Muhammad Habib Ullah. Sentimental analysis of covid-19 vaccine tweets using bert+ nbsvm. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 238–247. Springer, 2022.

[80] Areeba Umair, Elio Masciari, and Muhammad Habib Ullah. Vaccine sentiment analysis using bert+ nbsvm and geo-spatial approaches. *The Journal of Supercomputing*, pages 1–31, 2023.

[81] Areeba Umair, Muhammad Shahzad Sarfraz, Muhammad Ahmad, Usman Habib, Muhammad Habib Ullah, and Manuel Mazzara. Spatiotemporal Analysis of Web News Archives for Crime Prediction. *applied sciences*, 2020.

[82] James Warren and Nathan Marz. *Big Data: Principles and best practices of scalable realtime data systems.* Manning Publications Co., New York, 2015.

[83] Who. Covid-19 vaccine tracker and landscape. *WHO.*, 2021.

[84] Treepop Wisanwanichthan and Mason Thammawichai. A double-layered hybrid approach for network intrusion detection system using combined naive bayes and svm. *IEEE Access*, 9:138432–138450, 2021.

[85] Xiaoling Xiang, Xuan Lu, Alex Halavanau, Jia Xue, Yihang Sun, Patrick Ho Lam Lai, and Zhenke Wu. Modern Senicide in the Face of a Pandemic: An Examination of Public Discourse and Sentiment About Older Adults and COVID-19 Using Machine Learning. *J. Gerontol. B. Psychol. Sci. Soc. Sci.*, 76(4):e190–e200, 2021.

[86] Jia Xue, Junxiang Chen, Chen Chen, Chengda Zheng, Sijia Li, and Tingshao Zhu. Public discourse and sentiment during the COVID 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter. *PLoS One*, 15(9 September):1–12, 2020.

[87] Naina Yadav and Anil Kumar Singh. Bi-directional Encoder Representation of Transformer model for Sequential Music Recommender System. *ACM Int. Conf. Proceeding Ser.*, pages 49–53, 2020.

[88] Muhsin Yesilada and Stephan Lewandowsky. A systematic review: The youtube recommender system and pathways to problematic content. 2021.

[89] Fei Yi, Zhiwen Yu, Huang Xu, and Bin Guo. Talents Recommendation with Multi-Aspect Preference Learning. *Green, Pervasive, Cloud Comput.*, 11204:409–423, 2018.

[90] Hui Yin, Shuiqiao Yang, and Jianxin Li. *Detecting Topic and Sentiment Dynamics Due to COVID-19 Pandemic Using Social Media*, volume 12447 LNAI. Springer International Publishing, 2020.

[91] Beitong Zhou, Cheng Cheng, Guijun Ma, and Yong Zhang. Remaining Useful Life Prediction of Lithium-ion Battery based on Attention Mechanism with Positional Encoding. *IOP Conf. Ser. Mater. Sci. Eng.*, 895(1):0–9, 2020.

[92] Ziyu Zhou, Fang'Ai Liu, and Qianqian Wang. R-Transformer network based on position and self-attention mechanism for aspect-level sentiment classification. *IEEE Access*, 7:127754–127764, 2019.

[93] Bangren Zhu, Xinqi Zheng, Haiyan Liu, Jiayang Li, and Peipei Wang. Analysis of spatiotemporal characteristics of big data on social media sentiment with COVID-19 epidemic topics. *Chaos, Solitons and Fractals*, 140:110123, 2020.