Chapter 1

1. Introduction

1.1. Background Details

Sentiment Analysis is the practice of gathering and analyzing data that reflects individuals' emotions, reviews, and thoughts. It is often referred to as opinion mining because it extracts crucial insights from people's opinions. This analytical process utilizes a combination of deep learning techniques, machine learning techniques, statistical models, and Natural Language Processing (NLP) to extract features from large datasets. Sentiment Analysis can be conducted at different levels, including document, phrase, and sentence. At the document level, the entire document is summarized, and its sentiment is determined as positive, negative, or neutral. In phrase level analysis, the polarities of individual phrases within a sentence are assessed. At the sentence level, each sentence is classified into a specific sentiment category. Sentiment Analysis finds application in a wide range of fields and industries, showcasing its diverse applications.

Sentiment Analysis plays a significant role in generating opinions for individuals on social media platforms by analyzing the text-based expressions of their feelings and thoughts. It is important to note that Sentiment Analysis is domain-specific, meaning that the results obtained from analyzing one domain cannot be directly applied to another domain. In various real-life scenarios, Sentiment Analysis is employed to gather reviews and feedback. Twitter, for instance, serves as a microblogging platform where users can write and read short messages known as tweets. Considering the vast amount of data cumulated on Twitter, Sentiment Analysis becomes invaluable. The data obtained from Twitter is typically unstructured and written in natural language. Twitter Sentiment Analysis involves the process of accessing tweets related to a specific topic and utilizing different machine learning algorithms to predict the sentiment of these tweets as positive, negative, or neutral.

1.2. Introduction to Python

For this thesis, Python, a high-level dynamic programming language, was utilized. Specifically, Python version 3.11.3 was chosen due to its maturity, versatility, and robustness. Python is an interpreted language, which allows for rapid testing and debugging without the need for compilation. Furthermore, Python benefits from a wide range of open-source libraries and a large community of users, making it an excellent choice for various applications.

Python stands out as a programming language that is both simple and powerful. It is an interpreted and dynamic language that excels in processing natural language data, particularly spoken English, through libraries like NLTK (Natural Language Toolkit). While considering alternatives such as 'R' and 'MATLAB', which are also high-level programming languages known for their user-friendly interfaces, it becomes apparent that they do not provide the same level of flexibility and freedom that Python offers. Python's versatility and extensive libraries make it an excellent choice for natural language processing tasks.

1.3. Introduction to Tweepy

Tweepy is a widely used open-source Python package that provides a convenient means of accessing the Twitter API. It offers a range of classes and methods that represent Twitter's models and API endpoints. Tweepy simplifies the interaction with the Twitter API by seamlessly managing various implementation details, including:

- Data encoding and decoding
- HTTP requests
- Results pagination
- OAuth authentication
- Rate limits

- Streams

If Tweepy were not utilized, developers would have to handle low-level details related to HTTP requests, data serialization, authentication, and rate limits when accessing the Twitter API. Managing these aspects directly could be time-consuming and error-prone, requiring manual implementation and maintenance of these functionalities. However, Tweepy simplifies the process by providing a convenient and abstracted interface, allowing developers to focus on building the desired functionality without getting bogged down in the intricacies of API interactions. By leveraging Tweepy, developers can save time, reduce the risk of errors, and concentrate on implementing the core features and logic of their applications.

When comparing Twitter to other social media platforms, Twitter stands out as an effective platform for sharing informative messages during the COVID-19 pandemic. Active Twitter users often provide multiple insightful pieces of information regarding various aspects of the disease. These include details about the location and travel history of patients, the number of cases recovered, suspected, and confirmed, as well as the symptoms experienced by patients, such as body pains, running nose, headache, fever, and cold.

In order to distinguish between relevant and irrelevant content, COVID-19-related tweets are labeled as 'informative' tweets. These tweets contribute valuable information and insights regarding the disease. Conversely, tweets from users that do not provide meaningful information or contribute to the understanding of COVID-19 are labeled as 'uninformative' tweets.

This labeling process helps filter and focus on the tweets that offer significant and reliable information about the ongoing pandemic. By leveraging the informative tweets on Twitter, users can gather valuable insights and stay informed about the latest developments related to COVID-19. The major contributions of this study are:

- In order to enhance user sentiment prediction, the collected real-time Twitter data undergoes a process of data pre-processing. This involves eliminating various elements from the tweets, such as special characters, punctuations, numbers, repeated words, non-English characters, hashtag symbols, unnecessary spaces, tabs, and newlines. These steps are taken to clean and standardize the data, making it more suitable for accurate sentiment analysis. By removing these elements, the focus is directed towards the textual content of the tweets, which is crucial for effective sentiment prediction. Further, the exploratory investigation: keyword trend investigation and topic modeling are carried out for a better understanding of the collected data. In addition, the feature extraction is performed using TF-IDF, pre-trained Word2Vec, and fast text embedding for extracting the discriminative feature vectors from the pre-processed data.

- The ensemble classifier, composed of Gated Recurrent Units (GRU) and Capsule Networks (Caps Net), takes in the extracted discriminative feature vectors to classify people's sentiments into categories such as fear, joy, sadness, and anger. One challenge in training deep neural networks is the vanishing gradient problem. To mitigate this issue, the proposed ensemble classification model incorporates an activation function. Specifically, it employs a linear interpolation between the current candidate and prior candidates as the activation function. This choice of activation function helps address the vanishing gradient problem by ensuring a smooth gradient flow during the training process. By using this interpolation technique, the ensemble classifier enhances the overall performance of sentiment classification based on the extracted feature vectors.

- In this scenario, the effectiveness of the ensemble-based deep learning model is evaluated using various performance metrics such as f-score, accuracy, recall, and precision. These metrics provide a comprehensive assessment of the model's performance in sentiment classification.

The proposed ensemble-based deep learning model is compared to existing models like support vector machine (SVM) and logistic regression. The evaluation demonstrates that the ensemble model outperforms these existing models in terms of the evaluated performance metrics.

The ensemble model's superior performance can be attributed to the combination of GRU and Caps Net, which effectively capture the complex patterns and relationships in the data. By leveraging the strengths of both models, the ensemble classifier achieves better accuracy, recall, precision, and overall f-score compared to the other models.

The evaluation results validate the efficacy of the proposed ensemble-based deep learning model as an improved approach for sentiment classification tasks, showcasing its potential for enhancing sentiment analysis compared to traditional models.

f-score = 2*(precision*recall)/(precision*recall)

Accuracy = number of correct predictions.

For binary classifications, accuracy can also be calculated in term of positives and negatives as follows:

Accuracy = TP+TN/TP+TN+FP+FN.

Recall = TP/(TP+FN)

Precision = TP/TP+FP

In previous studies, the techniques for detecting events on Twitter were categorized, and their limitations were highlighted. However, these studies did not offer many solutions to address the issues associated with current approaches. To overcome these challenges and leverage the capability of learning text sequences and identifying relationships between words or phrases, the Contextual BiLSTM-CRF (CBC) model is employed in our suggested method for sentiment analysis.

The CBC model not only enhances the semantic understanding of tweets but also improves the efficiency of the learning model by capturing proper volatility patterns in the data. By incorporating the CBC model, the sentiment analysis process can yield better overall performance on datasets related to COVID-19.

To optimize the performance of the CBC model, a firefly optimization approach is proposed to tune its hyperparameters. This optimization technique helps recover the overall performance of the model, ensuring its effectiveness in sentiment analysis tasks. Furthermore, state-of-the-art ensemble and machine learning-based approaches are compared to the proposed model to evaluate its performance against alternative methods.

The suggested approach aims to address the shortcomings of existing event detection techniques on Twitter and improve sentiment analysis by employing the CBC model and optimizing its hyperparameters. By comparing it with other contemporary approaches, the proposed model demonstrates its potential in achieving enhanced performance and accuracy in sentiment analysis tasks.

1.4. Goal of the thesis

Dealing with large datasets can indeed be a challenge, but deep learning models offer a promising solution by enabling efficient classification and delivering more accurate results using various classifiers. In the context of this thesis, the objective is to conduct sentiment analysis on the topic of COVID-19. To achieve this, public opinions regarding the pandemic are collected from Twitter, and then classified into positive or negative sentiments using techniques such as LSTM (Long Short-Term Memory), CNN (Convolutional Neural Network), and other deep learning models.

By leveraging these models, the thesis aims to extract valuable insights from the data and gain a deeper understanding of people's reviews and opinions concerning COVID-19. The sentiment analysis results will shed light on the prevailing sentiment among individuals, providing valuable information about public perception and sentiment towards the pandemic.

Overall, the utilization of deep learning models in this thesis enables more accurate sentiment classification, helping to uncover valuable insights from the large dataset of Twitter data.

To accomplish this objective, a module is developed that enables live sentiment analysis. This module allows users to obtain real-time insights into the sentiment trend of any live trending topic, represented by two sentiment categories: positive and negative. These sentiments are visualized in live graphs, providing a dynamic representation of the evolving sentiment around the topic.

To ensure the accuracy and reliability of the module, various deep learning classifiers can be employed. These classifiers are utilized to evaluate the performance of the sentiment analysis module and assess the accuracy of the sentiment predictions. By comparing the module's results with the ground truth or known sentiments, the accuracy and reliability of the sentiment analysis module can be verified.

The integration of deep learning classifiers enhances the robustness and effectiveness of the module, improving its ability to accurately capture and classify sentiments in real-time. This live sentimental analysis module offers a valuable tool for monitoring public sentiment, tracking trends, and gaining insights into the sentiment dynamics surrounding different topics or events.

1.5. Needs of Sentiment Analysis

1.5.1. Industry Evolution

In industry, the focus is often on extracting relevant and useful information from a vast amount of data, rather than dealing with the entire unstructured dataset. Sentiment analysis plays a crucial role in this context by extracting important features from the data that are specifically relevant to the industry's objectives.

By performing sentiment analysis, industries can gain valuable insights that can be used to provide value and cater to the needs and preferences of their target audience. This applies to a wide range of industries, including restaurants, entertainment, hospitality, mobile customer services, retail, and travel.

For example, sentiment analysis can help restaurants gauge customer satisfaction and identify areas for improvement based on customer reviews. In the entertainment industry, sentiment analysis can help understand audience reactions to movies, TV shows, or music releases, aiding in decision-making and marketing strategies. In the hospitality sector, sentiment analysis can provide insights into guest experiences, helping hotels and resorts enhance their services.

Similarly, mobile customer services can leverage sentiment analysis to understand customer sentiment and address any issues or concerns promptly. Retail companies can analyze customer reviews and sentiments to optimize product offerings and improve customer experiences. In the travel industry, sentiment analysis can help identify popular destinations and preferences, assisting in personalized recommendations and targeted marketing.

Overall, sentiment analysis offers significant opportunities for industries to gain valuable insights, enhance customer satisfaction, and make informed business decisions, ultimately driving value and success.

1.5.2. Research Demand

The growth of sentiment analysis (SA) is driven by the increasing demand for research and advancements in evaluation, appraisal, opinion analysis, and their classification. Current solutions in sentiment analysis and opinion mining are rapidly evolving, particularly in reducing the amount of human effort required to classify and analyze comments.

Researchers are continually developing innovative approaches and techniques to automate and streamline the sentiment analysis process. This involves leveraging advancements in natural language processing (NLP), machine learning, and other related fields within computer science.

By reducing the need for manual classification and analysis, sentiment analysis solutions save time and resources while improving efficiency. This allows for the processing of large volumes of data, such as customer feedback, social media comments, and product reviews, in a more timely and cost-effective manner.

The research in sentiment analysis is built upon well-established disciplines within computer science, such as NLP, machine learning, data mining, and information retrieval. These disciplines provide the foundation for developing algorithms, models, and techniques that enable automated sentiment analysis and opinion mining.

As the field of sentiment analysis continues to evolve, research efforts are focused on enhancing the accuracy, scalability, and adaptability of

sentiment analysis methods. This ongoing research and development contribute to the growth and advancement of sentiment analysis as a valuable tool for understanding and classifying opinions in various domains, from customer feedback to social media analytics and beyond.

1.5.3. Decision Making

To obtain relevant information, a specific methodology is required for analyzing data and generating useful results. Performing regular surveys can be challenging, highlighting the importance of data analysis to identify the top products based on user opinions, reviews, and advice. These reviews and opinions are valuable in aiding individuals with important decision-making processes, particularly in research and business domains.

1.5.4. Understanding Contextual

With the increasing complexity of human language, it has become challenging for machines to fully comprehend the nuances, slangs, misspellings, and cultural variations inherent in human communication. Consequently, there arises a necessity for systems that facilitate better understanding and communication between human language and machine language. Such systems aim to bridge the gap and enable machines to interpret and respond effectively to the intricacies of human language.

1.5.5. Internet Marketing

Indeed, one significant factor contributing to the growing demand for sentiment analysis is the increasing use of internet marketing by businesses and organizations. With the rise of digital platforms, companies now actively monitor user opinions and sentiments about their brand, products, or events through blogs and social media posts. In this context, sentiment analysis serves as a valuable tool for marketing.

By analyzing the sentiments expressed by users, businesses gain insights into customer perceptions, preferences, and satisfaction levels. This Information helps them gauge the overall sentiment surrounding their brand and products, identify areas of improvement, and tailor their marketing strategies accordingly. Positive sentiments can be leveraged to reinforce brand loyalty and attract new customers, while negative sentiments can be addressed promptly to mitigate potential reputational risks.

Sentiment analysis enables businesses to understand the impact of their marketing efforts, assess the effectiveness of campaigns, and make data-driven decisions to enhance customer experiences. By leveraging sentiment analysis as a marketing tool, organizations can refine their messaging, improve customer engagement, and build stronger brand-consumer relationships.

Overall, sentiment analysis not only provides valuable insights into user opinions but also serves as a strategic resource for marketing departments, helping businesses stay attuned to customer sentiments and adapt their strategies accordingly.

1.6. Applications of Sentiment Analysis

Sentiment analysis plays a crucial role in numerous applications within the domain of Natural Language Processing (NLP). The growing significance of sentiment analysis has resulted in a high demand for social network data, as it offers valuable insights into user sentiments and opinions. Many companies have already embraced sentiment analysis to enhance their processes and operations.

By leveraging sentiment analysis, businesses can gain a deeper understanding of customer feedback, preferences, and trends. This enables them to make data-driven decisions, improve customer satisfaction, and refine their products or services. Sentiment analysis also helps companies monitor their brand reputation, identify emerging issues, and address customer concerns promptly.

In addition to business applications, sentiment analysis finds utility in various other domains. It is employed in social listening to understand public opinion on social and political issues. Sentiment analysis is also utilized in market research to assess consumer sentiment and gauge the success of marketing campaigns. Furthermore, sentiment analysis has applications in customer support, brand management, public opinion analysis, and even in the healthcare sector for analyzing patient feedback.

Overall, sentiment analysis has emerged as a valuable tool in NLP, offering insights and actionable information for businesses across diverse industries. The adoption of sentiment analysis enables companies to make informed decisions, improve customer experiences, and stay ahead in today's competitive market.

1.6.1. Word of Mouth (WOM)

Sentiment Analysis becomes crucial in the context of Word of Mouth (WOM), where people share information and opinions about businesses, services, and products. Online platforms like review blogs and social networking sites provide a wealth of opinions, making decision-making easier. Sentiment Analysis helps analyze and understand these opinions, aiding in informed decision-making.

1.6.2. Voice of Voters

Political parties allocate a significant portion of their budget towards campaign efforts and influencing voters. By understanding the opinions,

reviews, and suggestions of the people, politicians can enhance the effectiveness of their campaigns. Sentiment analysis plays a crucial role in this process, benefiting not only political parties but also news analysts. It enables a deeper understanding of public sentiment, aiding politicians in shaping their messaging and strategies. Similar techniques have already been employed by the British and American administrations to gain insights from public opinion.

1.6.3. Online Commerce

Many ecommerce websites have a feedback policy in place, allowing users and customers to provide their opinions and experiences. This feedback encompasses various aspects, including service quality, product features, and suggestions. Companies collect and compile these details and reviews, often converting them into a geographical format. This format helps visualize and analyze the data, providing insights into user experiences and opinions. The data collected from these ecommerce websites also reflects the current techniques employed by online commerce platforms, ensuring that the information stays up to date.

1.6.4. Voice of the Market (VOM)

When a company is preparing to launch a new product, customers are interested in accessing ratings, reviews, and detailed descriptions to make informed decisions. Sentiment analysis plays a significant role in this process by analyzing marketing and advertising data, enabling companies to develop effective strategies for promoting their product. By utilizing sentiment analysis, companies can gain valuable insights into customer sentiments and preferences, helping them tailor their marketing efforts to target specific audiences. This empowers customers to make informed choices by considering the sentiments expressed by

others, ultimately allowing them to select the best product among the available options.

1.6.5. Government

Sentiment analysis assists administrations in providing public services by generating fair results that analyze both positive and negative aspects. It proves useful in decision making, recruitment, taxation, and evaluating social strategies. Similar techniques contribute to citizen-centric government models, prioritizing services, and preferences. Applying sentiment analysis in multilingual countries like India, where content is generated in multiple languages, presents an interesting challenge that can be addressed for effective analysis of citizen sentiments.

1.7. Deep learning

Deep learning is a powerful approach in sentiment analysis that utilizes neural networks and other techniques to analyze and comprehend the sentiment or emotional content of text data. Its goal is to ascertain the subjective attitude or opinion expressed in text, classifying it as positive, negative, or neutral. Deep learning models in sentiment analysis harness the capabilities of artificial neural networks (ANN) to learn and extract significant representations from textual data.

The various deep learning methods are:

- Artificial neural networks (ANN) are the foundation of deep learning. ANNs (Artificial Neural Networks) are composed of interconnected nodes, known as neurons, organized in layers. These neurons receive input signals, perform mathematical operations, and generate outputs. ANNs can have multiple hidden layers, allowing them to learn intricate patterns and representations. By stacking these layers, ANNs can capture hierarchical relationships and extract high-level features from the input data. This ability to learn complex patterns makes ANNs well-

suited for tasks such as sentiment analysis, where the understanding of nuanced and intricate language patterns is crucial.

- Convolutional neural networks (CNN) are designed to process data with a grid-like structure, such as images. CNNs (Convolutional Neural Networks) use specialized layers called convolutional layers to capture spatial patterns in data. They excel in tasks like image classification, object detection, and image segmentation. These layers enable CNNs to automatically extract meaningful features from the input, allowing them to understand complex patterns and structures in images. This makes CNNs particularly effective for analyzing visual data. Recurrent neural networks (RNN) are suited for sequential data analysis, such as natural language processing and speech recognition. They utilize feedback connections, allowing them to maintain information from previous inputs. This enables RNNs to capture temporal dependencies and model sequences effectively.

- Long Short-Term Memory (LSTM) networks are a type of RNN that address the “vanishing gradient” problem. They introduce memory cells that can selectively retain or forget information over time, making them well-suited for tasks involving long-term dependencies.

- Autoencoders are unsupervised learning models that aim to learn efficient representation of data. They consist of ana encoder network that compresses the input data into a latent space representation, and a decoder network that reconstructs the original data from the latent representation. Autoencoder are used for tasks such as dimensionality reduce, anomaly detection etc.

- Reinforcement Learning (RL) involves an agent that learns to interact with an environment to maximize a reward signal. Deep RL algorithm

combines deep neural network with RL techniques, enabling agents to learn complex policies in domains such as game playing, robotics, etc.

- Bag-of-words (BOW) represents as a collection of unique words or terms and their frequencies. It disregards the order and structure of text. Each tweet is transformed into a vector, where each dimension corresponds to a word, and value represents its frequencies or occurrence.

- Term Frequency-Inverse Document Frequency (TF-IDF) considers not only the frequency of words in a tweet but also their importance in overall dataset. It calculates a weight for each word in a tweet based on its frequency within the tweet and its rarity across the entire data.

- Word embeddings captures semantic relationships between words by mapping them to a continuous vector space. Techniques like Word2Vec, GloVe, and fast text are commonly used to generate word embeddings.

Chapter 2

2. Literature Review

The majority of research on sentiment analysis relies heavily on machine learning and deep learning algorithms. These algorithms aim to determine whether a given text expresses a favorable or unfavorable sentiment and to ascertain the polarity of the text. In this article, we will present an overview of several research studies that contribute to a deeper understanding of this subject matter.

The COVID-19 pandemic has caused many deaths and is a major global health threat. Social media and financial markets have reacted to the pandemic, showing different aspects such as mortality rates, contagion factors, and timing in different countries. During the lockdowns, social media played an important role in sharing information and allowing people to express their feelings about the pandemic. It is important to analyze people's reactions on Twitter by focusing on common words related to the epidemic.

We can utilize natural language processing (NLP) methods such as textual feature extraction, sentiment analysis, and word cloud visualizations to delve into the vast and intricate discussions surrounding this topic. In our research, we employed deep learning classifiers to perform sentiment analysis on tweets related to COVID-19.

2.1. Jelodar Hamed, Yongli Wang, Rita Orji, Hucheng Huang, et al.

In a prior study, the authors employed a recurrent deep-learning method to classify the sentiment of microblogging messages related to COVID-19. They conducted text analysis on Reddit to identify themes associated with COVID-19, whereas our research centered around analyzing data from Twitter. Our approach involved a combination of deep learning and traditional machine learning techniques to

classify the tones of tweets, while they solely relied on long short-term memory (LSTM) as a deep learning approach.

2.2. Gupta et al.

Gupta et al. introduced an emotion care approach to analyze real-time COVID-19 tweets, using multi-modal text data. They initially transformed the collected tweets into lowercase strings and removed punctuation, special characters (except AZ, a-z), user mentions, retweets, links, and stop-words. After data cleaning, tokenization was applied to break sentences into words. They then developed a multi-modal vector by identifying frequently used words in the tweets using the Term Frequency-Inverse Document Frequency (TF-IDF) technique. Finally, the authors classified the emotions into eight scales: trust, surprise, sadness, joy, fear, anticipation, anger, and disgust.

2.3. Majumder et al.

Majumder et al. conducted a study where they collected Twitter data from across India. The collected data was preprocessed by converting it to lowercase, removing hyperlinks, punctuations, and abbreviations of retweets. Label encoding was applied to convert the data labels (negative, neutral, and positive tweets) into numeric form (0, 1, and 2), facilitating the data classification process. Lemmatization and Text Blob were then employed, followed by classification using Support Vector Machine (SVM) and logistic regression models to classify the emotions expressed in the tweets.

2.4. S. Imran et al.

In their study, S. Imran et al. focused on analyzing the sentiments of individuals concerning the COVID-19 lockdown measures implemented by various countries. To accomplish this, they utilized a deep learning approach known as Long Short-Term Memory (LSTM) network. The LSTM network enabled them to estimate emotions and sentiment polarities expressed in people's

tweets, providing valuable insights into the overall sentiment surrounding the lockdown actions taken by different countries.

By leveraging the LSTM network, the researchers were able to capture the temporal dependencies and context within the tweet data, allowing for a more nuanced analysis of people's sentiments. This deep learning approach facilitated the estimation of emotions and sentiment polarities, providing a comprehensive understanding of how individuals felt about the COVID-19 lockdown measures.

Through their study, S. Imran et al. shed light on the sentiment landscape surrounding the lockdown actions implemented by various countries. By employing the LSTM network, they were able to extract valuable insights and contribute to the broader understanding of public sentiments during the COVID-19 pandemic.

2.5. Naseem et al.

Naseem et al. collected Twitter data from the COVID-Senti dataset and used Text Blob for labeling emotional sentiments as neutral, negative, or positive. They applied data cleaning techniques to preprocess the Twitter data, considering its noisy, informal, short, and unstructured nature. Improved word vector and hybrid word ranking methods were employed to incorporate tweet context and sentiments for sentiment analysis. Finally, sentiment classification was performed using various deep and machine learning techniques such as LSTM networks, SVM, decision trees, Naïve Bayes, CNN, and random forests.

2.6. Samrat et al.

In their research, Samrat et al opted to employ the supervised K-Nearest Neighbor (KNN) technique as a means to classify the sentiments expressed by

individuals regarding COVID-19 vaccination. The K-Nearest Neighbor algorithm, often referred to as KNN or k-NN, is a non-parametric and supervised learning classifier. It operates by utilizing the proximity or distance measures between data points to make predictions or classifications.

In the context of sentiment analysis, the KNN algorithm works by assigning sentiments to a particular data point based on the sentiments of its nearest neighbors. The "k" in KNN represents the number of neighboring data points considered when making the classification. The algorithm calculates the proximity of the given data point to its k nearest neighbors and assigns the sentiment label that is most prevalent among those neighbors.

By using the KNN technique, the study by Samrat et al aimed to leverage the proximity-based approach in sentiment classification. This allowed them to predict the sentiments of individuals concerning COVID-19 vaccination based on the sentiments expressed by similar neighboring data points.

2.7. Ortega et al.

Ortega et al. introduced a three-step technique for Twitter sentiment analysis, which includes pre-processing, polarity detection, and rule-based classification. The polarity detection and rule-based classification steps rely on SentiWordNet and WordNet. The approach shows promising results when evaluated on the SemEval-2013 dataset. However, its effectiveness has not been compared to other existing technologies for tweet sentiment analysis.

2.8. Saif et al.

In their research, Saif et al. presented SentiCircles, a novel lexicon-based approach designed for sentiment analysis. SentiCircles focuses on analyzing word patterns in various contexts, updating word polarity, and assigning sentiment scores. The

researchers conducted evaluations using three distinct datasets and discovered that SentiCircles consistently outperformed other methods such as SentiWordNet and MPQA-based approaches.

Furthermore, the study acknowledges the significant role of deep learning in the field of natural language processing (NLP). Deep learning techniques have made noteworthy contributions to sentiment analysis and other NLP tasks. These methods leverage neural networks with multiple layers to learn intricate patterns and representations from textual data, enabling them to capture complex relationships and improve the accuracy of sentiment analysis models.

By highlighting the superior performance of SentiCircles and acknowledging the impact of deep learning in NLP, Saif et al.'s research contributes to the advancement of sentiment analysis techniques and emphasizes the importance of leveraging innovative approaches and technologies in the field.

2.9. Chandra et al.

In their work, Chandra et al. put forward a framework for sentiment analysis based on deep learning techniques. They utilized Long Short-Term Memory (LSTM) and Bidirectional LSTM models along with GloVe word embeddings for language modeling purposes. Furthermore, they conducted a comparative analysis between LSTM and Bidirectional LSTM models using the Bidirectional Encoder Representations from Transformers (BERT) model.

By employing these deep learning models and embeddings, Chandra et al. aimed to enhance the understanding of sentiment analysis in the context of the pandemic, specifically focusing on sentiments expressed in India. They evaluated the performance of different models and identified the best-performing one to carry out sentiment analysis during the pandemic.

The utilization of LSTM and Bidirectional LSTM models, coupled with GloVe word embeddings and comparison with BERT, demonstrates the authors' exploration of various deep learning techniques for sentiment analysis. This approach allows for a more robust and comprehensive analysis of sentiments during the pandemic, providing valuable insights into the prevailing sentiments and emotions within the Indian context.

Chapter 3

3. Methodology

Sentiment analysis is a valuable tool used to uncover and understand people's attitudes, feelings, and opinions by analyzing their comments on social media platforms. With the rapid advancements in machine learning and deep learning techniques, sentiment analysis on Twitter has gained significant traction among researchers. In this particular research, the primary objective is to delve into people's opinions and sentiments surrounding the topic of COVID-19.

The experimental results obtained from this study provide profound insights into the sentiments and viewpoints expressed by individuals in relation to COVID-19. To ensure the collection of relevant data, a careful selection of keywords was made, including terms such as Quarantine, social distancing, lockdown, coronavirus, corona, Covid-19, corona outbreak, pandemic, stay home, and coronavirus outbreak. These keywords were instrumental in capturing a wide range of discussions and perspectives on Twitter.

The proposed framework for Twitter sentiment analysis comprises five distinct phases, each serving a crucial purpose in the overall process. Firstly, Twitter data collection involves gathering a substantial volume of tweets that are relevant to the research topic. Next, the collected data undergoes thorough pre-processing, which involves cleaning and transforming the data to ensure its quality and consistency. The exploratory analysis phase aims to gain valuable insights and uncover patterns and trends within the dataset.

Feature extraction, another critical phase, involves identifying and extracting meaningful features from the tweet data. These features serve as valuable inputs for the subsequent sentiment prediction phase. By employing various machine learning

and deep learning techniques, sentiment prediction models are built to accurately classify and predict the sentiment associated with each tweet.

Overall, this comprehensive framework allows researchers to gain a deep understanding of people's opinions, emotions, and attitudes towards COVID-19 by effectively analyzing and interpreting the extensive Twitter data available.

3.1. Twitter data collection

In this research, the data of users are collected for COVID-19. The generated dataset consists of tweets which are extracted from Twitter API. After data collection, every tweet is annotated as ‘sad,’ ‘joy,’ ‘fear’, and ‘anger’. In this research, the Text Blob tool is utilized to label the emotional sentiments into ‘sad,’ ‘joy,’ ‘fear,’ and ‘anger.’

The TextBlob tool assesses the sentiment of a sentence by calculating a polarity score ranging from -1 to 1. Sentiments are classified as "joy" if the polarity score exceeds 0.1. Similarly, sentiments are classified as "fear," "sad," and "anger" if the polarity scores fall within the ranges of -0.1 to -0.3, -0.4 to -0.7, and -0.8 to -1, respectively. The mathematical representation of polarity score estimation is provided below.

$$L_{Ti} = \{\text{Joy} \quad P > 0.1$$
$$\text{Fear} \quad -0.1 \geq P \leq -0.3$$
$$\text{Sad} \quad -0.4 \geq P \leq -0.7$$
$$\text{Anger} \quad -0.8 \geq P \leq -1\}$$

Where, P_i indicates polarity of tweet T_i.

Pseudocode of tweets labeling

Input:

Unlabeled tweet T_{IJ}

Output:

Labeled tweet T_L

Calculate:

Sad: $T_{Sad} = [];$

Joy: $T_{Joy} = [];$

Fear: $T_{Fear} = [];$

Anger: $T_{Anger} = [];$

Steps:

For t is English in T(t) do:

If(t):

Perform labeling using Text Blob tool

If ($-0.4 \geq$ Polarity o f t ≤ -0.7):

Labelled as ‘sad’

If ($-0.1 \geq$ Polarity o f t ≤ -0.3):

Labelled as ‘fear’

If ($-0.8 \geq$ Polarity o f t ≤ -1):

Labelled as ‘anger’

Else

Labelled as ‘joy’

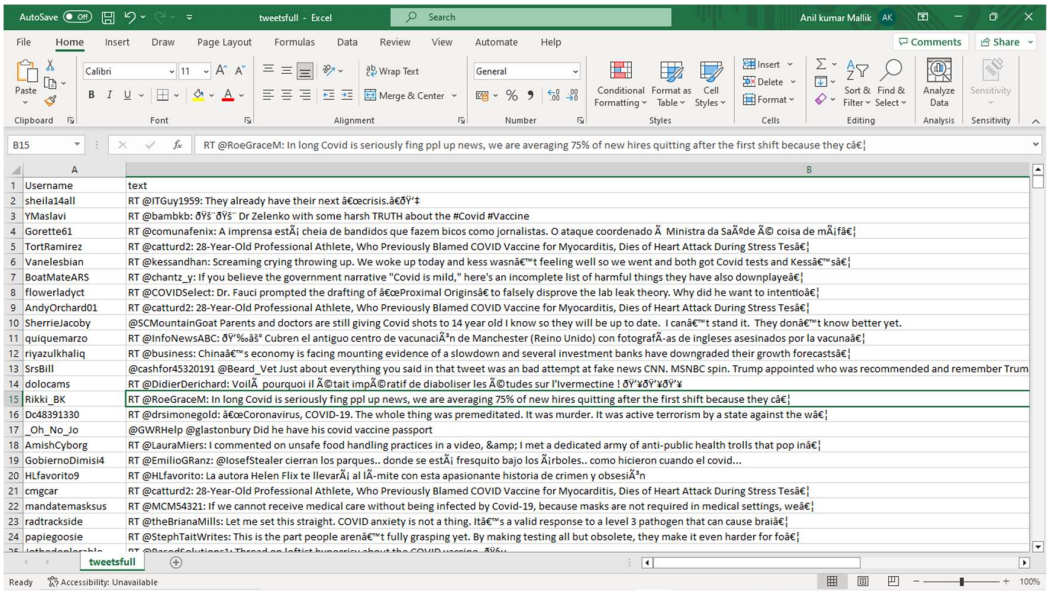
End For

Output:

Labelled tweets: $T_L = [T_{Joy}, T_{Sad}, T_{Fear}, T_{Anger}]$.

```
# Write the header row
csv_writer.writerow(["Username", "Text", "Retweets", "Favorites", "Sentiment"])

# Iterate over the tweets and write them to the CSV file
for tweet in tweets:
    username = tweet.user.screen_name
    text = tweet.full_text
    retweets = tweet.retweet_count
    favorites = tweet.favorite_count

    # Perform sentiment analysis
    analysis = TextBlob(text)
    sentiment = analysis.sentiment.polarity

    # Write the tweet data and sentiment to the CSV file
    csv_writer.writerow([username, text, retweets, favorites])

    # Check rate limit status
    remaining_requests = api.rate_limit_status()["resources"]["search"]["/search/tweets"]["remaining"]
    if remaining_requests == 0:
        # Sleep for the specified time
        sleep_time = api.rate_limit_status()["resources"]["search"]["/search/tweets"]["reset"] - time.time()
        print(f"Rate limit reached. Sleeping for {sleep_time} seconds...")
        time.sleep(sleep_time)

# Print a success message
print(f"Tweets retrieved and saved to {csv_file_path}")
```

```
Tweets retrieved and saved to tweets.csv
```



	A	B
1	Username	text
2	sheila14all	RT @ITGuy1959: They already have their next â€œcrisis.â€ðŸ‘‡
3	YMaslavi	RT @bambkb: ðŸš¨ Dr Zelenko with some harsh TRUTH about the #Covid #Vaccine
4	Gorette61	RT @comunafenix: A imprensa estÃ¡ cheia de bandidos que fazem bicos como jornalistas. O ataque coordenado Ã Ministra da SaÃºde Ã© coisa de mÃ¡fiâ€¦
5	TortRamirez	RT @catturd2: 28-Year-Old Professional Athlete, Who Previously Blamed COVID Vaccine for Myocarditis, Dies of Heart Attack During Stress Tesâ€¦
6	Vanelesbian	RT @kessandhan: Screaming crying throwing up. We woke up today and kess wasnâ€™t feeling well so we went and both got Covid tests and Kessâ€™sâ€¦
7	BoatMateARS	RT @chantz_y: If you believe the government narrative "Covid is mild," here's an incomplete list of harmful things they have also downplayeâ€¦
8	flowerladyct	RT @COVIDSelect: Dr. Fauci prompted the drafting of â€œProximal Originsâ€ to falsely disprove the lab leak theory. Why did he want to intentioâ€¦
9	AndyOrchard01	RT @catturd2: 28-Year-Old Professional Athlete, Who Previously Blamed COVID Vaccine for Myocarditis, Dies of Heart Attack During Stress Tesâ€¦
10	SherrieJacoby	@SCMountainGoat Parents and doctors are still giving Covid shots to 14 year old I know so they will be up to date. I canâ€™t stand it. They donâ€™t know better yet.
11	quiquemarzo	RT @InfoNewsABC: ðŸ‘‰ Cubren el antiguo centro de vacunaciÃ³n de Manchester (Reino Unido) con fotografÃas de ingleses asesinados por la vacunaâ€¦
12	riyazulkhaliq	RT @business: Chinaâ€™s economy is facing mounting evidence of a slowdown and several investment banks have downgraded their growth forecastsâ€¦
13	SrsBill	@cashfor45320191 @Beard_Vet Just about everything you said in that tweet was an bad attempt at fake news CNN. MSNBC spin. Trump appointed who was recommended and remember Trum
14	dolocams	RT @DidierDerichard: VoilÃ pourquoi il Ã©tait impÃ©ratif de diaboliser les Ã©tudes sur l'Ivermectine ! ðŸ‘¥ðŸ‘¥ðŸ‘¥
15	Rikki_BK	RT @RoeGraceM: In long Covid is seriously fing ppl up news, we are averaging 75% of new hires quitting after the first shift because they câ€¦
16	Dc48391330	RT @drsimonegold: â€œCoronavirus, COVID-19. The whole thing was premeditated. It was murder. It was active terrorism by a state against the wâ€¦
17	_Oh_No_Jo	@GWRHelp @glastonbury Did he have his covid vaccine passport
18	AmishCyborg	RT @LauraMiers: I commented on unsafe food handling practices in a video, & I met a dedicated army of anti-public health trolls that pop inâ€¦
19	GobiernoDimisi4	RT @EmilioGRanz: @IosefStealer cierran los parques.. donde se estÃ¡ fresquito bajo los Ã¡rboles.. como hicieron cuando el covid...
20	HLfavorito9	RT @HLfavorito: La autora Helen Flix te llevarÃ¡ al lÃmite con esta apasionante historia de crimen y obsesiÃ³n
21	cmgcar	RT @catturd2: 28-Year-Old Professional Athlete, Who Previously Blamed COVID Vaccine for Myocarditis, Dies of Heart Attack During Stress Tesâ€¦
22	mandatemasksus	RT @MCM54321: If we cannot receive medical care without being infected by Covid-19, because masks are not required in medical settings, weâ€¦
23	radtrackside	RT @theBrianaMills: Let me set this straight. COVID anxiety is not a thing. Itâ€™s a valid response to a level 3 pathogen that can cause braiâ€¦
24	papiegoosie	RT @StephTaitWrites: This is the part people arenâ€™t fully grasping yet. By making testing all but obsolete, they make it even harder for foâ€¦

3.2. Data pre-processing

After the data collection or the tweets collection, the quality of the raw labelled data is enhanced by performing the pre-processing operations.

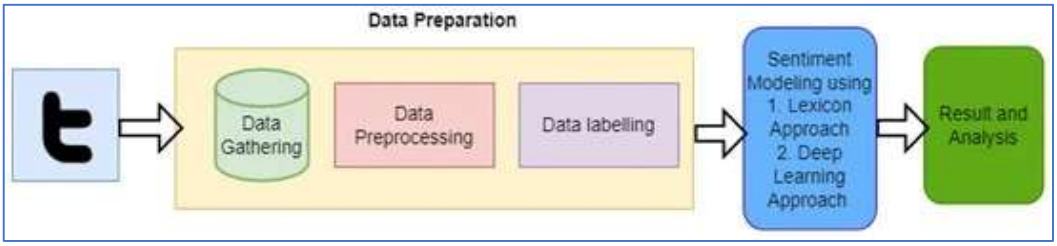
- Eliminate numbers, punctuations, and special characters from the dataset, where it majorly will not improve the prediction performance.

- Eliminate repeated words. For instance: “sooooo boring” is converted as “so boring.”

- Eliminate non-English characters because this study mainly concentrated on the analysis of information in English language.

- Eliminate hashtag symbols (#china, #lockdown, #Wuhan, etc.), uniform resource locators, and @users from the tweets, because it will not contribute to analyzing the messages.

- Eliminate un-necessary newlines, tabs, and spaces from the tweets.



(Data pre-processing)

- The emoji are converted into short textual description using python emoji2 library.

3.3. Exploratory investigation

Once the data has undergone pre-processing, the next step is to conduct an exploratory investigation to gain a more comprehensive understanding of the datasets. The exploratory investigation consists of two important steps: keyword trend investigation and topic modeling.

The first step, keyword trend investigation, involves analyzing the trends and patterns of the selected keywords within the dataset. By examining the frequency and distribution of these keywords over time, researchers can identify the prominence and popularity of specific terms related to COVID-19. This analysis provides valuable insights into the key topics and themes that dominate discussions on social media platforms like Twitter.

The second step in the exploratory investigation is topic modeling. This technique aims to automatically identify and extract latent topics or themes present in the dataset. By applying advanced machine learning algorithms, topic modeling enables researchers to uncover underlying patterns and themes within the data without prior knowledge or explicit labeling. This process assists in organizing and categorizing the tweets based on shared topics or concepts, contributing to a deeper understanding of the prevalent discussions and sentiments surrounding COVID-19.

Overall, the exploratory investigation following the data pre-processing stage plays a crucial role in gaining a more comprehensive view of the datasets. It involves analyzing keyword trends to identify the most prominent terms related to COVID-19 and utilizing topic modeling to uncover latent topics and themes within the data. These steps provide researchers with valuable insights for further analysis and interpretation of people's opinions and sentiments towards COVID-19.

- Keyword trend investigation

A keyword trend analysis is performed on the pre-processed Twitter data to identify the most frequently mentioned words. The common topics that people discuss include social distancing, staying at home, coronavirus cases, coronavirus pandemic, COVID outbreak, and the crisis caused by the coronavirus.

- Topic modeling

  LDA is a statistical topic modeling technique used to analyze topic distribution in the dataset. It represents documents as a mixture of topics, where each topic consists of a group of words. LDA captures topics from weighted features and classifies each tweet based on these concepts. It assigns probability distributions to describe the topics in tweets using Dirichlet distributions.

3.4. Feature extraction

After conducting exploratory analysis on the data, the next step in the sentiment analysis process involves feature extraction using CNN and LSTM techniques.

The CNN (Convolutional Neural Network) technique is employed to capture local patterns and features in the textual data. Convolutional layers are applied to the input data, allowing the network to automatically learn important features through convolutions and pooling operations.

Simultaneously, the LSTM (Long Short-Term Memory) technique is utilized to capture sequential dependencies and long-term dependencies in the text. LSTM networks are capable of retaining and utilizing context information by incorporating memory cells and gating mechanisms.

By employing both CNN and LSTM techniques for feature extraction, the sentiment analysis model can effectively capture both local patterns and sequential dependencies within the text. This combined approach leverages the strengths of both architectures to extract meaningful and informative features that contribute to accurate sentiment analysis.

3.5. Sentiment Prediction

The Twitter sentiment analysis utilizes a classifier that combines the power of both CNN (Convolutional Neural Network) and LSTM (Long Short-Term Memory) models. This combination allows for the utilization of both convolutional and recurrent neural network architectures in analyzing the sentiment of Twitter data.

The CNN component of the classifier is particularly effective in capturing local patterns and features within the text of each tweet. It achieves this by leveraging convolutional layers to extract meaningful features and learn representations from the input data.

On the other hand, the LSTM component, being a type of recurrent neural network, excels at capturing sequential and long-term dependencies within the tweet data. It processes the input sequence of words or characters and maintains an internal memory to retain contextual information throughout the analysis.

By integrating both CNN and LSTM architectures, the classifier aims to leverage the respective strengths of each model. The CNN is adept at extracting local features, while the LSTM is capable of capturing temporal dependencies. This combined approach enables the classifier to effectively analyze the sentiment expressed in Twitter data and classify it into appropriate sentiment categories, such as positive, negative, or neutral.

In summary, the integration of CNN and LSTM models in the classifier allows for a comprehensive analysis of Twitter sentiment. The CNN

component captures local patterns, while the LSTM component captures long-term dependencies, resulting in a powerful framework for sentiment analysis on Twitter data.

Chapter 4

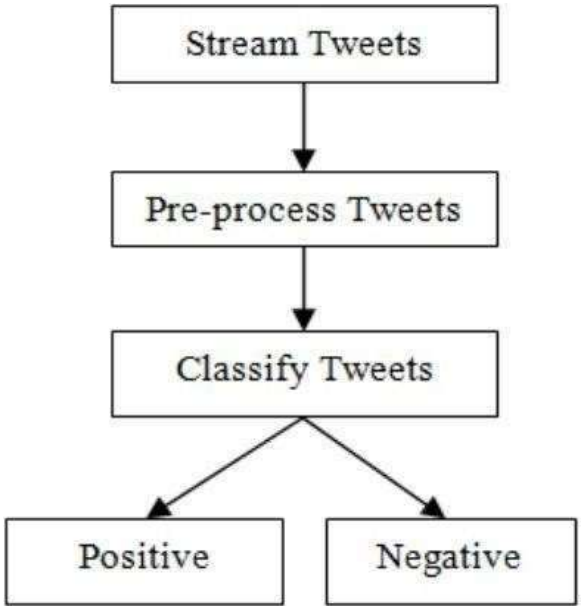
4. Implementation

Gathering data for analysis is a multifaceted undertaking that demands meticulous decision-making. In the scope of our thesis, we embarked on the creation of datasets that were meticulously tailored to cater to the requirements of training, testing, and carrying out Twitter sentiment analysis. This chapter serves as a focal point for comprehending the intricacies associated with data collection, storage, processing, and classification. However, prior to delving into the nitty-gritty of these processes and the diverse datasets we utilized, it is of utmost importance to provide a comprehensive overview of the architecture we propose for our research.

4.1. Proposed Architecture

As our goal is to achieve sentiment analysis for data provided from Twitter. We are going to build a classifier which consists of different machine learning classifiers. Once our classifier is ready and trained, we are going to follow the steps as shown in Figure.

(Process to classify tweets using build classifier)



Step-1: First we are going to stream tweets in our build classifier with the help of tweepy library in python.

Step-2: Then we pre-process these tweets, so that they can be fit for mining and feature extraction.

Step-3: After pre-processing we pass this data in our trained classifier, which then classify them into positive or negative class based on trained results.

Since, Twitter is our source of data for analysis. We are going to stream the tweets from twitter in our database. For this we are going to use Twitter Application.

- Twitter API (Application Programming Interface)

  Twitter provides two types of APIs for collecting tweets: REST API and Streaming API. The REST API allows for short-term connections and limited data retrieval at a time. On the other hand, the Streaming API provides real-time access to tweets and allows for long-term connections. In our analysis, we utilize the Streaming API for collecting many tweets. This API allows us to maintain a long-lived connection and has no limitations on data rate.

4.2. Twitter Data Collection

To utilize the Twitter API, it is necessary to have a Twitter account. Twitter provides a platform that allows us to access data from Twitter accounts and use it for our own purposes. In the context of this thesis, our objective is to analyze the sentiment of tweets related to COVID-19. Therefore, we should focus on collecting tweets specifically related to this topic.

Python, being a powerful programming language, offers numerous services through various libraries. Tweepy is an open-source Python library that enables communication between Python and Twitter's API, allowing us to collect data

from Twitter for use in our program. With the help of Tweepy, we can easily access and retrieve tweets for further analysis and sentiment classification.

4.3. Data Storage

After retrieving the data from the Twitter API, the next step is to store the data in a suitable format for sentiment analysis. In this case, a .csv format is chosen due to its ability to organize data with multiple fields. CSV files use commas to separate each field, making it easy to access specific fields, such as the text of the tweets. Additionally, CSV files offer fast read/write times compared to other formats.

To manage the collected data effectively, separate directories are created to store tweets related to different COVID-19 cases. This allows for better organization and retrieval of specific data sets when needed.

However, before applying the data to a sentiment classifier, it is essential to preprocess the stored data. The data retrieved from the API may not be suitable for analysis directly and requires preprocessing.

4.4. Data Pre-Processing

Data obtained from Twitter often requires preprocessing before it can be effectively used for feature extraction. Tweets typically contain various elements such as usernames, empty spaces, special characters, stop words, emoticons, abbreviations, hashtags, timestamps, URLs, and more. To prepare the data for mining, preprocessing techniques are applied using deep learning functions.

The preprocessing steps involve extracting the main message from the tweet and removing unwanted elements such as empty spaces, stop words

(commonly used words with little semantic meaning), hashtags, repeated words, URLs, etc. Emoticons and abbreviations are also replaced with their corresponding meanings to ensure consistency in the data. For example, expressions like :-) or =D could be replaced with "happy" or "laugh."

Cleaning the Twitter data is crucial as it eliminates unnecessary syntactic features, allowing the data to be represented solely in terms of words that are more relevant for analysis. The preprocessed tweets are then provided to the classifier for further analysis and sentiment classification.

We create a code in Python in which we define a function which will be used to obtain processed tweet. This code is used to achieve the following functions:

∑ remove quotes - provides the user to remove quotes from the text.

∑ remove @ - provides choice of removing the @ symbol, removing the @ along with the username, or replace the @ and the username with a word 'AT_USER' and add it to stop words.

∑ remove URL (Uniform resource locator) - provides choices of removing URLs or replacing them with 'URL' word and add it to stop words.

∑ remove RT (Re-Tweet) - removes the word RT from tweets.

∑ remove Emoticons - remove emoticons from tweets and replace them with their specific meaning.

∑ remove duplicates – remove all repeating words from text so that there will be no duplicates.

∑ remove # - removes the hash tag class.

∑ remove stop words – remove all stop words like a, he, the, etc. which provides no meaning for classification.

Removed and modified content

CONTENT	ACTION
Punctuation (! ,? , . ” : ;)	Removed
#word	Removed #word
@any_user	Remove @any_user or replaced it with “AT_USER” and then added in stop words.
Uppercase characters	Lowercase all content
URLs and web links Remove URLs or re	placed with “URL” and then added stop words
Number	Removed
Word not starting with alphabets	Removed
All Word	Stemmed all word. (Converted into the simple form)
Stop words	Removed
Emoticons	Replaced with respective meaning
White spaces	Removed

Sample cleaned data.

Raw data	Clean data
@jackstenhouse69, I liked it, in me opinion it def is :)	Liked, opinion, def
:(\u201c@EW: How awful. Police: Driver kills 2, injures 23 at #SXSW http://t.co/8GmFiOuZbS\u201d	Sad, awful, police, driver, kills, Injuries

Chapter 5

5. Result and Discussion

In this application, the performance of the proposed ensemble-based deep learning model is evaluated using the Python environment. The evaluation of the model is conducted for each country, considering metrics such as prediction accuracy, recall, MCC ,precision, and F-score.

Precision is defined as the ratio of correctly identified positive instances to the total instances predicted as positive. Recall, on the other hand, represents the ratio of correctly identified positive instances to the total actual positive instances. The F-score is a weighted harmonic mean of precision and recall.

The mathematical expressions for prediction accuracy, recall, precision, and F-score are as follows: [Mathematical expressions can be provided based on the specific formulas used in the evaluation.].

$$\text{Accuracy} = \left(\frac{TP + TN}{TP + TN + FP + FN} \right) * 100$$

$$\text{Recall} = \frac{TP}{TP + FN \times 100}$$

$$\text{Precision} = \frac{TP}{TP + FP \times 100}$$

$$F - score = \left(\frac{2TP}{FP + 2TP + FN} \right) *100$$

Where, True Positive (TP) represents that the number of informative tweets is predicted correctly, False Positive (FP) denotes that the number of informative tweets is predicted incorrectly, True Negative (TN) indicates that the number of uninformative tweets is predicted correctly and False Negative (FN) represents that the number of uninformative tweets is predicted incorrectly. Following the performance analysis on Accuracy, Recall, MCC, Precision, F1-score of different methods.

	Feature Extraction	Classifier	Accuracy	Recall	MCC	Precision	F-Score
0	CNN	CNN	0.955501	0.955501	809	0.957421	0.955799
1	CNN	CNN	0.939431	0.939431	0.864133	0.948926	0.941572
2	LSTM	LSTM	0.943140	0.943140	0.868066	0.948494	0.943675
3	CNN	CNN	0.940668	0.940667	0.863621	0.945883	0.941736
4	LSTM	LSTM	0.955501	0.955501	0.896575	0.958745	0.955695

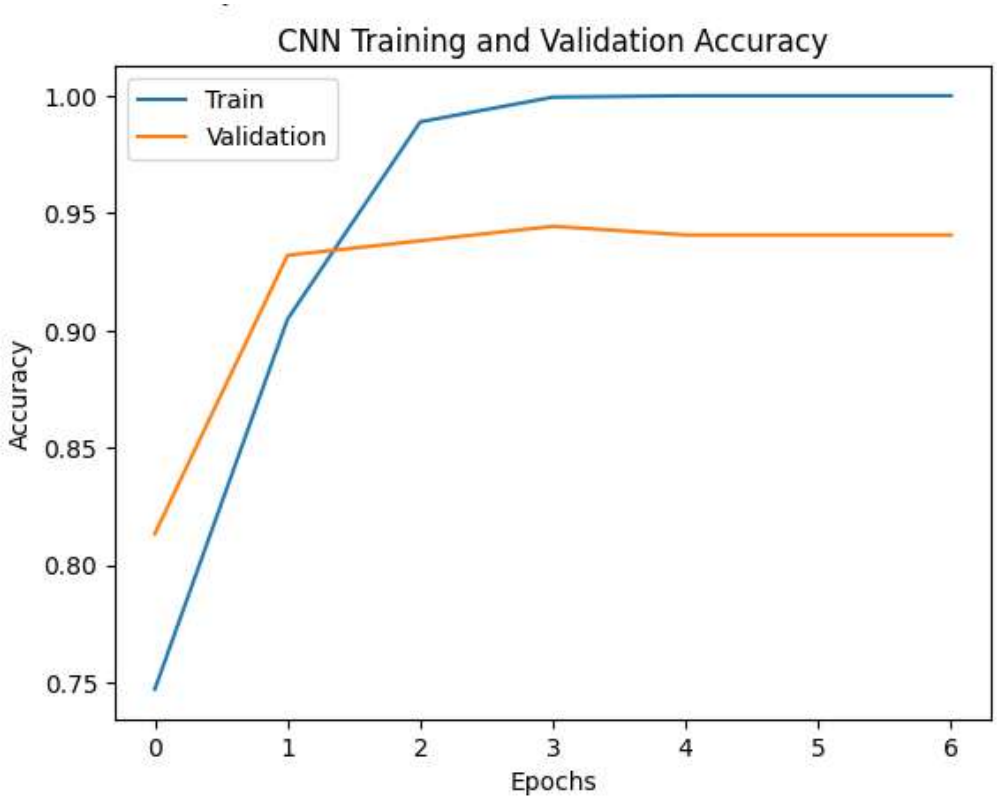
(Performance Analysis)

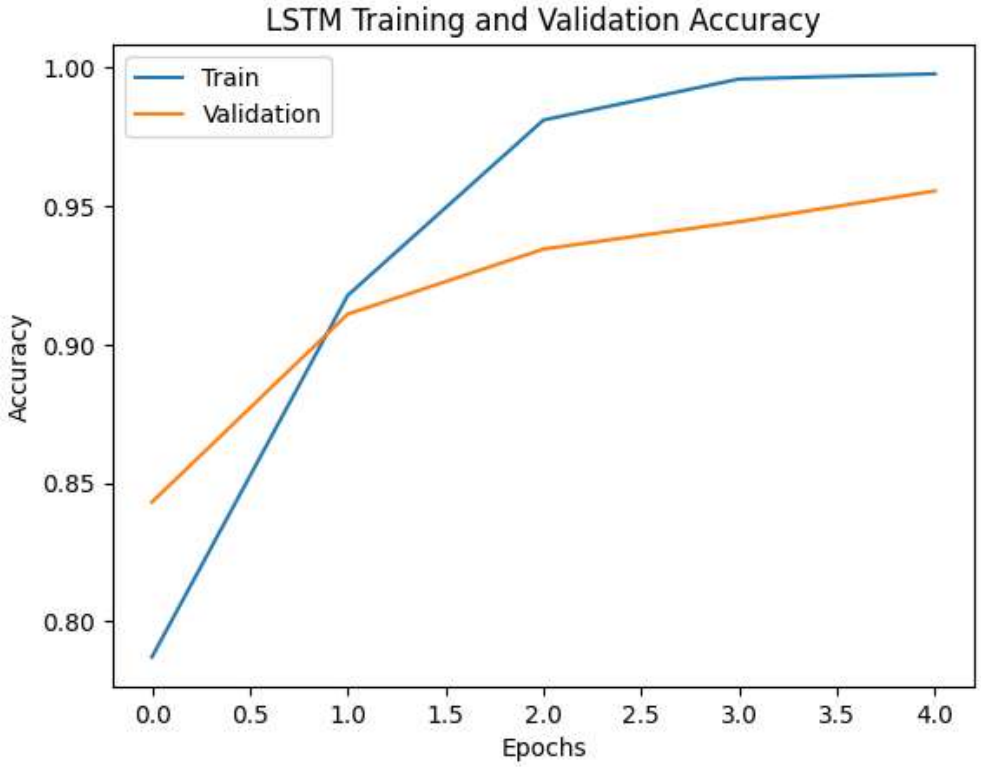
In this quantitative study, a total of 4500 tweets were collected from people. These tweets were used for training and testing the proposed ensemble-based deep learning model as well as other classifiers with different feature extraction techniques. The collected tweets serve as the dataset for evaluating the performance of the models and analyzing the sentiment expressed in the tweets.

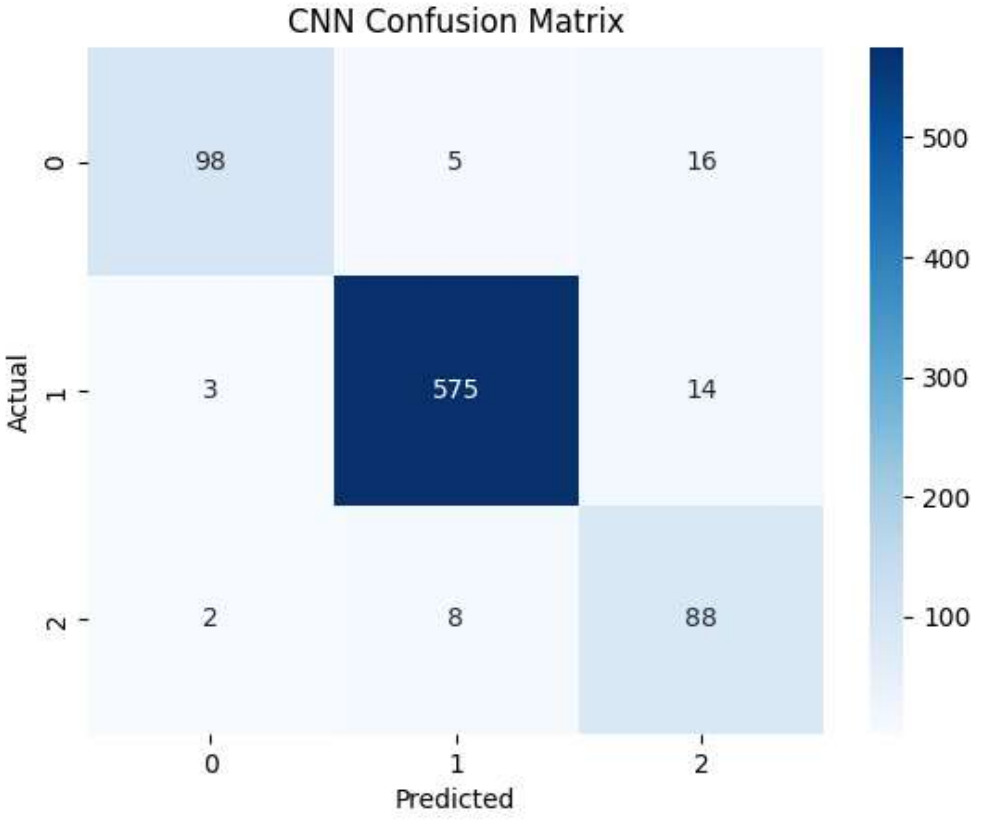
5.1. Tweets collection and Analysis

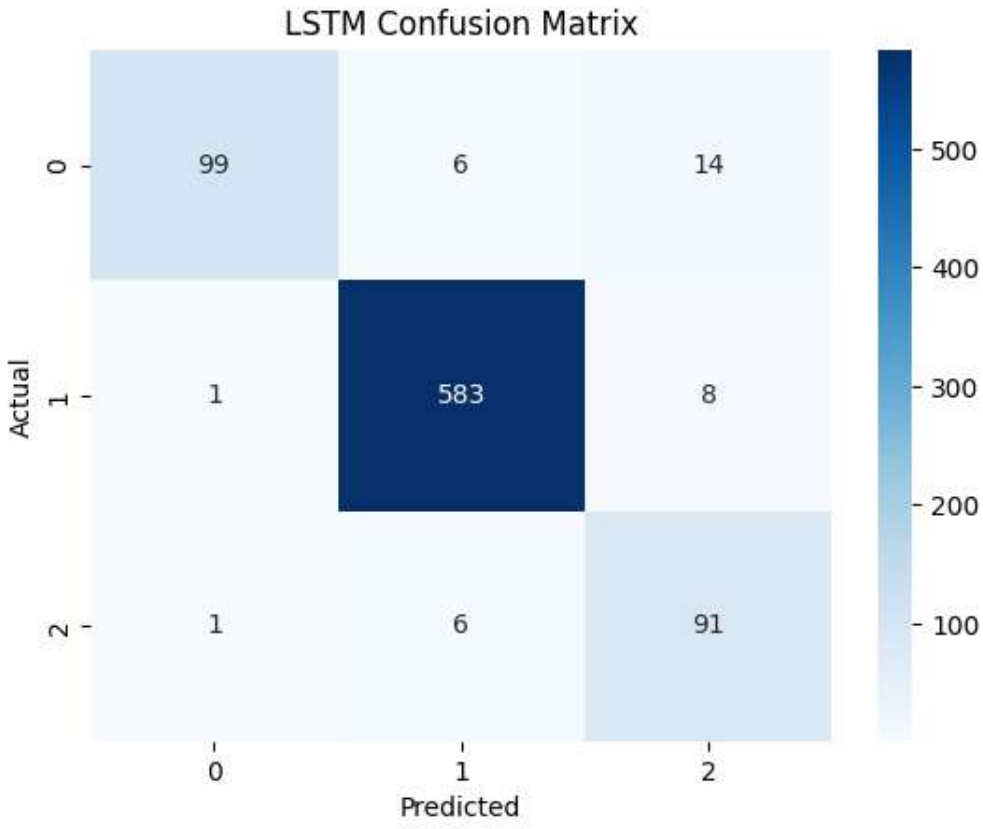
Tweets were collected by using various hashtags and prompts related to COVID-19 at different time periods through the Twitter API. All the collected

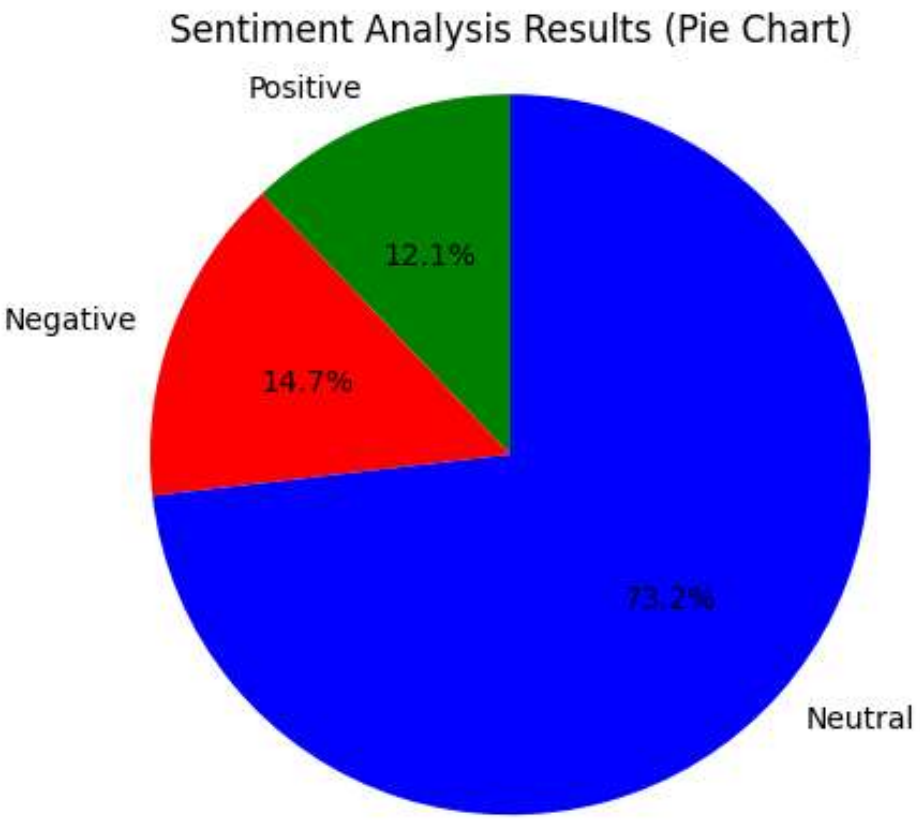
tweets were saved in a single dataset. After collecting enough tweets, further steps were taken, including training, and testing the model to calculate the test loss and test accuracy. Following that, a pie chart and a bar graph were created to visualize the percentages of positive, negative, and neutral tweets in the dataset.

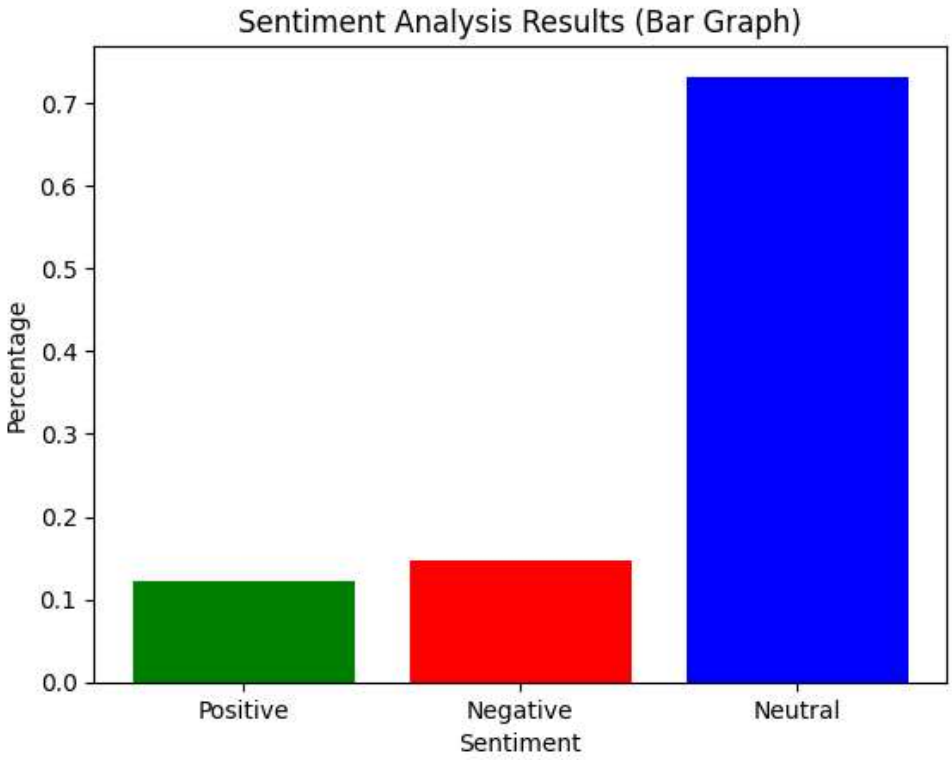

CNN Training and Validation Accuracy
Train
Validation
1.00
0.95
0.90
0.85
0.80
0.75
Accuracy
0
1
2
3
4
5
6
Epochs


LSTM Training and Validation Accuracy
Train
Validation
1.00
0.95
0.90
0.85
0.80
Accuracy
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Epochs


CNN Confusion Matrix
Actual
Predicted
0
1
2
98
5
16
3
575
14
2
8
88
500
400
300
200
100


LSTM Confusion Matrix
Actual
Predicted
0
1
2
99
6
14
1
583
8
1
6
91
500
400
300
200
100


Sentiment Analysis Results (Pie Chart)
Positive
12.1%
Negative
14.7%
73.2%
Neutral


Sentiment Analysis Results (Bar Graph)
Percentage
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Positive
Negative
Neutral
Sentiment

5.2. Discussion:

The spread of pandemics creates vulnerability and anxiety among the global population, making effective risk communication crucial for implementing preventive strategies. By bridging the knowledge gap and providing genuine information, risk communication aims to understand public behavior and combat the pandemic. Sentiment analysis helps understand people's emotions during such events. In the case of COVID-19, sentiment analysis of tweets revealed positive sentiments expressing gratitude for community efforts and support for frontline workers, along with negative opinions towards some individuals. Joyful statements encouraged doctors, while support for infection control measures and following health guidelines was also evident. In India, sentiment analysis during the COVID-19 lockdown highlighted interesting results, emphasizing the need for further validation and experimentation. Deep learning models and classification techniques can be employed for autonomous sentiment classification of COVID-19. However, this thesis is limited by relying on a single word ("fear") from US citizens. Other studies have identified major topics related to healthcare problems and highlighted negative sentiments among Chinese youth. In contrast, this study focused on four emotions of Indian netizens, encompassing both positive and negative sentiments.

- Challenges in Sentiment Analysis

  Sentiment Analysis is a very challenging task. Following are some of the challenges faced in Sentiment Analysis of Twitter.

  a. Identifying subjective parts of text: Identifying subjective parts of text in sentiment analysis is challenging because the same word can be subjective in one case and objective in another. This variability makes it difficult to accurately determine the subjective portions of text.

b. Domain dependence: Domain dependence is another challenge in sentiment analysis. The meaning of a sentence or phrase can vary depending on the specific domain or context in which it is used. For example, the word "unpredictable" may have a positive connotation in the domain of movies or dramas, but it can have a negative opinion when used in the context of a vehicle's steering. Considering the domain-specific understanding of words and phrases is crucial for accurate sentiment analysis.

c. Sarcasm Detection: Sarcastic sentences use positive words in a unique way to express a negative opinion about a target.

d. Explicit Negation of sentiment: Sentiment can be negated in many ways as opposed to using simple no, not, never, etc. It is difficult to identify such negation.

e. Order dependence: Discourse Structure analysis is essential for Sentiment Analysis/Opinion Mining. For example, A is better than B, conveys the exact opposite opinion from, B is better than A.

f. Entity Recognition: There is a need to separate out the text about a specific entity and then analyze sentiment towards it.

g. Handling noisy text: social media posts, product reviews, or online comments often contain noisy or informal language, misspellings, abbreviations, slangs, or grammatical errors. Handling such noise and understanding sentiments in these texts can be difficult.

h. Lack of labeled data: Building accurate sentiment analysis models required labeled data for training. However, obtaining large, diverse, and accurately labeled datasets can be time-consuming, expensive, and labor-intensive.

Chapter 6

6. Conclusion and future scope

Sentiment analysis serves the purpose of identifying and understanding people's opinions, perspectives, and emotional states, which can range from positive to negative. Adjectives play a significant role in extracting sentiment as they provide descriptive information. However, when adjectives and adverbs are used together, determining sentiment becomes more challenging.

In the proposed system for sentiment analysis of tweets, the initial step involves collecting user posts from Twitter. The collected data then undergoes preprocessing to enhance its quality and remove noise or irrelevant information. Following this, feature extraction and exploratory analysis are performed to extract relevant features from the preprocessed data. These extracted feature vectors are then fed into an ensemble classifier, which combines multiple classifiers for sentiment prediction.

The experimental results of the study demonstrate that the proposed ensemble-based deep learning model outperforms individual feature extraction techniques and classifiers in terms of accuracy, recall, precision, and f-score. The model achieved a CNN test loss of 0.1765 and a CNN test accuracy of 0.94 and LSTM test loss of 0.268 and LSTM test accuracy of 0.955 when classifying people's sentiments in Twitter sentiment analysis. However, it is worth noting that the proposed method exhibits high computational complexity, particularly when dealing with large feature lengths.

In summary, the proposed system presents a comprehensive approach to sentiment analysis of tweets. It employs preprocessing, feature extraction, exploratory analysis, and an ensemble-based deep learning model for sentiment prediction. The experimental results validate the effectiveness of the proposed approach, although

its computational complexity may pose challenges when dealing with extensive feature lengths.

6.1. Future Scope

Indeed, sentiment analysis has garnered considerable attention in recent years owing to its extensive applicability across diverse industries. The ability to analyze and understand sentiment expressed in text data has proven invaluable in numerous domains such as marketing, customer service, brand reputation management, social media analytics, market research, and more.

In the realm of marketing, sentiment analysis enables businesses to gauge customer perceptions and attitudes towards their products or services. It aids in identifying customer preferences, gathering feedback, and monitoring brand sentiment in real-time. This information can be leveraged to make data-driven decisions and improve marketing strategies.

For customer service, sentiment analysis helps companies assess customer satisfaction levels by analyzing customer feedback and sentiment expressed in customer support interactions. It allows for proactive identification of potential issues, enabling prompt resolution and enhancing overall customer experience.

Sentiment analysis also plays a vital role in brand reputation management. By monitoring sentiment on social media platforms, companies can quickly identify and address any negative sentiment or potential PR crises. It enables proactive reputation management and facilitates timely interventions to maintain a positive brand image.

In market research, sentiment analysis provides valuable insights into consumer opinions and preferences. It helps researchers understand market trends, consumer behavior, and emerging patterns, enabling organizations to

make informed decisions regarding product development, pricing, and market positioning.

Furthermore, sentiment analysis has proven beneficial in political analysis, financial markets, healthcare, and other domains where understanding public sentiment is crucial for decision-making and forecasting.

Overall, sentiment analysis has emerged as a powerful tool with significant applications across various industries. Its ability to extract valuable insights from textual data, interpret sentiment, and understand human emotions has revolutionized decision-making processes and enhanced customer experiences in numerous sectors.

The future scope of sentiment analysis looks promising and includes the following areas:

a. Use of parser can be embedded into system to improve results.

b. A web-based application can be made for our work in future.

c. We can improve our system that can deal with sentences of multiple meanings.

d. We can also increase the classification categories so that we can get better results.

e. We can start work on multi languages like Hindi, Spanish, and Arabic to provide sentiment analysis to more local.

References

[1] Jelodar, H., Wang, Y., Orji, R., & Huang, S. (2020). Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics*, *24*(10), 2733-2742.

[2] Vedika Gupta, Nikita Jain, Piyush Katariya, Adarsh Kumar, Senthilkumar Mohan, Ali Ahmadian, Massimiliano Ferrara, An Emotion Care Model using Multimodal Textual Analysis on COVID-19, Chaos, Solitons & Fractals, Volume 144, 2021, 110708, ISSN 0960-0779.

[3] Majumder, S., Aich, A. and Das, S., 2021. Sentiment analysis of people during lockdown period of COVID-19 using SVM and logistic regression analysis. Available at SSRN 3801039

[4] A. S. Imran, S. M. Daudpota, Z. Kastrati and R. Batra, "Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on COVID-19 Related Tweets," in IEEE Access, vol. 8, pp. 181074-181090, 2020, doi: 10.1109/ACCESS.2020.3027350.

[5] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund and J. Kim, "COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis," in IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 1003-1015, Aug. 2021, doi: 10.1109/TCSS.2021.3051189.

[6] Muruganantham, P., Wibowo, S., Grandhi, S., Samrat, N. H., & Islam, N. (2022). A systematic literature review on crop yield prediction with deep learning and remote sensing. *Remote Sensing*, *14*(9), 1990.

[7] Pizarro-Ortega, C. I., Dioses-Salinas, D. C., Severini, M. D. F., López, A. D. F., Rimondino, G. N., Benson, N. U., ... & De-la-Torre, G. E. (2022). Degradation of plastics associated with the COVID-19 pandemic. *Marine pollution bulletin*, *176*, 113474.

[8] Jahrami, H. A., Alhaj, O. A., Humood, A. M., Alenezi, A. F., Fekih-Romdhane, F., AlRasheed, M. M., ... & Vitiello, M. V. (2022). Sleep disturbances during the COVID-19 pandemic: a systematic review, meta-analysis, and meta-regression. *Sleep medicine reviews*, *62*, 101591.

[9] Chandra, R., Krishna, A.: Covid-19 sentiment analysis via deep learning during the rise of novel cases. PLoS One. 16(8), e0255615 (2021).

[10] W. Zhai, Z.R. Peng, F. Yuan, Examine the effects of neighborhood equity on dis-aster situational awareness: harness machine learning and geotagged Twitter data, Int. J. Disaster Risk Reduct. 48 (2020) 101611.

[11] R. Mahajan, V. Mansotra, predicting geolocation of tweets: using combination of CNN and BiLSTM, Data Sci. Eng. 6 (4) (2021) 402–410.

[12] A. Arora, P. Chakraborty, M.P.S. Bhatia, P. Mittal, Role of emotion in excessive use of Twitter during COVID-19 imposed lockdown in India, J. Technol. Behav.Sci. 6 (2) (2021) 370–377.

[13] P. Gupta, S. Kumar, R.R. Suman, V. Kumar, Sentiment analysis of lockdown in India during COVID-19: a case study on twitter, IEEE Trans. Comput. Soc. Syst. (2020).
[14] A.S. Imran, S.M. Daudpota, Z. Kastrati, R. Batra, Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets, IEEE Access 8 (2020) 181074–181090.
[15] U. Naseem, I. Razzak, M. Khushi, P.W. Eklund, J. Kim, Covidsenti: a large-scale benchmark Twitter data set for COVID-19 sentiment analysis, IEEE Trans. Com- put. Soc. Syst. 8 (4) (2021) 1003–1015.
[16] F.J.M. Shamrat, S. Chakraborty, M.M. Imran, J.N. Muna, M.M. Billah, P. Das,M.O. Rahman, Sentiment analysis on Twitter tweets about COVID-19 vaccinesusing NLP and supervised KNN classification algorithm, Ind. J. Electric. Eng.Comput. Sci. 23 (1) (2021) 463–470.
[17] N. Chintalapudi, G. Battineni, F. Amenta, Sentimental analysis of COVID-19 tweets using deep learning models, Infect. Dis. Rep. 13 (2) (2021) 329–339.