Human Sentiment Analysis using Natural Language Processing

by

Md. Shamiul Haque Khan Shamya
19101208
Md. Shariar Hossain Fahim
19101367
Niaz Makhdum Khan
19101371
K.M Mahfuzul Monir
19101463
Mubasshir Hussain
19101552

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
School of Data and Sciences
Brac University
May 2023

Declaration

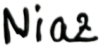
It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Md. Shamiul Haque Khan Shamya
19101208

Md. Shariar Hossain Fahim
19101367



Niaz Makhdum Khan
19101371

K.M Mahfuzul Monir
19101463

Mubasshir Hussain
19101552

Approval

The thesis/project titled "Human Sentiment Analysis using Natural Language Processing" submitted by

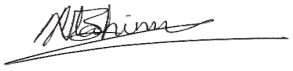
1. Md. Shamiul Haque Khan Shamya (19101208)
2. Md. Shariar Hossain Fahim (19101367)
3. Niaz Makhdum Khan (19101371)
4. K.M Mahfuzul Monir (19101463)
5. Mubasshir Hussain (19101552)

Of Spring, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on May 22, 2023.

Examining Committee:

Supervisor:

(Member)



Nabuat Zaman Nahim
Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr.Sadia Hamid Kazi
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Emotion is something that roams with us, the vibrant tapestry of our human existence. People are frequently so emotionally motivated that we may get a peek of their emotions simply by observing them, listening to them, or reading their text messages/social status. For this reason it is hard sometimes to detect someone's feelings. Therefore, we have devised a plan to examine and comprehend the sentiment of a person as determined via sentiment analysis which is a part of natural language processing (NLP). To determine various values from various data configurations, we used three hybrid models: CNN-LSTM, LSTM, and RNN. Six target variables made up the first smaller datasets we worked on, and then we moved on to some larger datasets. People today tend to frequently post their ideas on social media. They want the world to know about their suffering, hope, and joy. Our main objective is to build a program or project that uses sentiment analysis to analyze a person's writing and generate a report about his or her emotion by that. We also intend to acquire useful real-world datasets. In a contemporary society where people are so emotionally motivated, we aim to recognize their sentiment digitally via natural language processing where we can help a lot of people by that.

Keywords: Sentiment, CNN-LSTM, LSTM, RNN, Emotion, Machine Learning, NLP , Prediction, Token, Embedding

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.
Secondly, to our supervisor Nabuat Zaman Nahim sir for his kind support and advice in our work. He helped us whenever we needed help.
Thirdly, Jon Van Haaren and the whole judging panel of Machine Learning in Sports Analytics Conference 2015. Though our paper not accepted there, all the reviews they gave helped us a lot in our later works.
And finally to our parents without their throughout sup-port it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Background

We are living in the era of science where interaction between humans and technology has reached a new dimension. The modern world is entirely powered by artificial intelligence and machine learning. Where AI has fundamentally altered many concepts of livelihood by enabling AI to learn how to perform many tasks without even the assistance of a human. Similar to this, sentiment analysis, a branch of machine learning, can be used to observe and recognize human sentiments for artificial intelligence. Sentiment analysis, a machine learning technique, analyzes texts to determine if they have a positive, negative, or neutral tone. Machine learning technology can be taught to recognize sentiment automatically without human input by being fed examples of text expressing different emotions. Additionally, it is a field that seeks to teach computers how to comprehend human emotions.
Positive, neutral, and negative emotions equivalent to +1, 0 and -1, respectively, from a computer's perspective. Each sentence's keywords are grouped. In addition, it is subdivided into other sections, such as emotion recognition, aspect-based sentiment analysis, multilingual sentiment analysis, etc. Emotional detection is one of the most important components of sentiment analysis, and it is achieved through the use of lexicons (lists of words and the emotions they express) or complicated machine learning algorithms. In this emotion detection, Kman's six primary emotions (anger, disgust, fear, happiness, sorrow, and surprise) are employed to identify human emotions.
The significance of sentiment analysis cuts across many different fields. First off, it offers companies insightful data on client views, preferences, and levels of satisfaction. Organizations can identify advantages and disadvantages in their goods and services, make data-driven decisions for improvement, and improve overall customer experiences by analyzing sentiment from consumer feedback, reviews, and social media posts. Second, sentiment analysis is essential for managing brands and maintaining reputations. It's essential to keep up a positive brand image in the connected world of today. Businesses can proactively identify possible reputation hazards by watching and analyzing sentiment related to a brand or organization in real-time. This allows them to respond to customer issues quickly and maintain a favorable brand image. Third, market research and competition analysis greatly benefit from sentiment analysis. Companies may gain knowledge about new market demands, assess the competitive environment, and develop smart business plans to

stay ahead through tracking views regarding market trends and products, or services. A variety of strategies and procedures are used to achieve sentiment analysis. These models make use of innovative techniques like CNN-LSTM, a hybrid network design that combines the advantages of convolutional neural networks (CNNs) and long short-term memory (LSTM), which are recurrent neural networks. It is frequently used for processing geographically structured sequential data, including 2D picture or video data. The LSTM component records time dependence and long-term context, whereas the CNN component retrieves spatial information from the input data. By implementing this, the community may be able to avert some big issues like customer Insights, brand Monitoring, market Research, social Media and campaign Analysis and so on. Additionally, it can aid those who require support but are unable to find it in a moment of need. At the very least, it can bring a massive change to the betterment of human life.

1.2 Problem Statement

Sentiment or emotion is a vital component of human existence. Social media platforms are the most popular location for people to express their opinions with the world that show their individuality and their emotional response to a certain event and their current mental state. This has both beneficial and detrimental impacts. Frequently, there are a large number of posts that are uplifting, constructive, and hilarious; occasionally, there are a large number of posts that encourage people and support lawbreaking. These items have an effect on society, and individuals may be deceived by this information. Therefore, our objective is to identify such sentiments and, if possible, take further action. Natural Language processing is the key to achieving our objective.

Natural Language Processing, sometimes known as NLP, is an area of machine learning concerned with the computer and statistical modeling of human languages. The field of NLP is wide and intricate. To achieve our goal of evaluating and comprehending human emotion, we must overcome a number of limitations of this sophisticated technology.

To begin with, it is difficult to record the vocabularies and rules of human languages due to the fact that the interpretation of human language differs depending on the incident's context and the language's variety, ambiguity, and grammatical structure [12]. For instance, the term 'star' typically refers to an astronomical object, but it can also refer to a person's notoriety in a different context.

The following concepts are grammar and parser. Grammar and parser applications are extremely important in NLP since they are used to characterize a sentence in order to construct a computational model from it [1]. Every NLP system must have a grammar and a parser, which is essentially an algorithm that examines the sentence structure. According to reports, numerous NLP systems are based on CFG- Context Free Grammar. Here, the grammar G consists of three fundamental components: I non-terminals, such as NP: Noun Phrase and V: verb; (ii) terminals, such as Fahim, likes, and chocolate; and (iii) a collection of rewriting rules for terminals and non-terminals, known as syntactic rules and lexical rules. All of these components will be limited in quantity. The difficulty with the CFG is that it is not always adequate to adequately characterize a sentence. To illustrate, it necessitates a complicated informational package called feature-structure, which is

essentially a collection of attribute-value pairs. However, only the syntactic category adheres to this structure. Moreover, extremely intricate string-combining procedures and Additionally, tree-combining processes are required to understand a sentence's fundamental concepts. In addition to grammatical difficulties, parsing complexity is another factor to consider when designing CFG. The parser is not always adaptable. The issue with parsers is that it is extremely difficult to rate all the parses of a sentence without statistical data. Additionally, a parser is not necessarily adaptable, as it may be incapable of parsing a subset of sentences.
Moreover, data mining is an underlying issue in our research concerning the concept of collecting data from social networks. Data mining is a process used to retrieve information and data from digital sources [4]. However, it is not that straightforward. When discussing the retrieval of data from social platforms, an ethical question arises because it involves issues of privacy and security. Therefore, there may be a contradiction between the user's permission to use social media and the researcher's aim to utilize it for study purposes. The authors of [25] argue that the protection of personal data should be incorporated into privacy rights. In addition to individual privacy, consideration must also be given to group privacy. Regarding ethics, it is still controversial how much of a person's information may be collected.

1.3 Research Objective

Finding the precise feeling that someone on social media communicated is the major goal of the research. The six emotions that make up human sentiment that have been chosen for inquiry are basically all different. The six major emotions are joy, sadness, anger , disgust, fear and surprise. The objectives of this research are:

- Finding the best result using long short-term memory networks (LSTM), convolutional neural networks (CNN), and recurrent neural networks (RNN).
- Using grammar and parser correctly to achieve the optimum results.
- Verifying the accuracy of the analysis.
- Developing a better strategy for improving the model.
- Establishing a strong suggestive model for the after-work result.

Chapter 2

Literature Review

Sentiment Analysis is the study of human emotion that is expressed through speech, text, image, body language in a computational approach. It is a sub-field of natural language processing that has become a very popular research area now-a-days. It deals not only with research and analytical purposes but also marketing and business purposes as well. There are three main levels of sentiment analysis. These are: document level, sentence level and aspect level [10]. Document level sentiment analysis is used to extract sentiment from a whole document, whereas sentence level takes one particular sentence to identify sentiment within. For the aspect level, it detects sentiment with regard to characteristics or features of an entity person toward an event or occurrence. Researchers have developed a bunch of sophisticated algorithms for using sentiment analysis. An overall figure of how sentiment analysis is accomplished is given below: Along with the abundance in various sectors of today's world

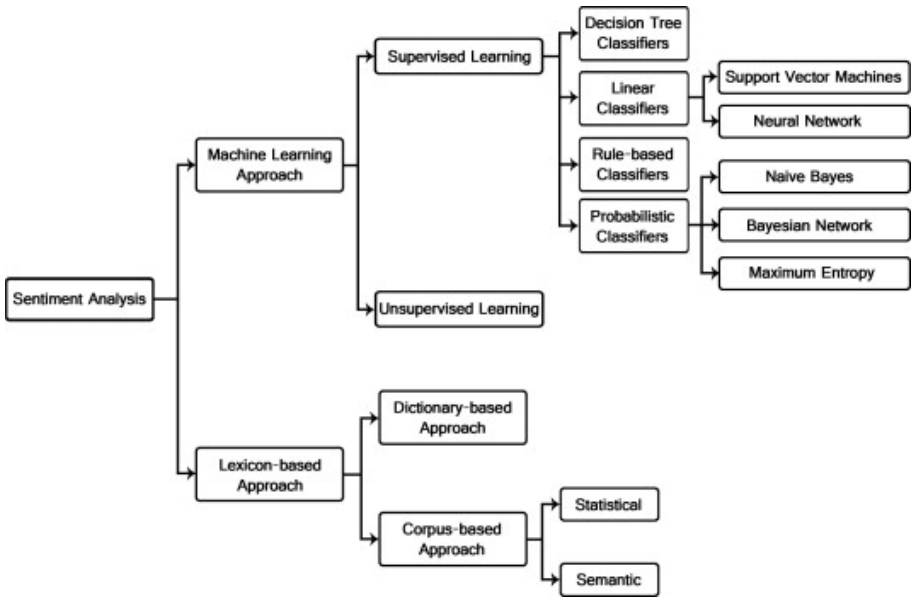


Figure 2.1: Approach to Sentiment Analysis [[40]]

of science, sentiment analysis comes with a few challenges as well. The paper carries out an empirical study to visualize the challenges of sentiment analysis and they illustrated it by means of two things: review structure and accuracy improvement [27]. First, they analyzed 41 papers and compared them to find out the challenges. The result showed that the relevant factor in this case is the orientation of the topic

domain and its features during sentiment analysis. Secondly, the comparison was conducted to mark the challenges in terms of applying the techniques and improving their accuracy. The study further showed that the most used techniques are parts of speech tagging and lexicon based approach. Two figures depicting the analysis are given below: A handful of machine learning algorithms are available today to help

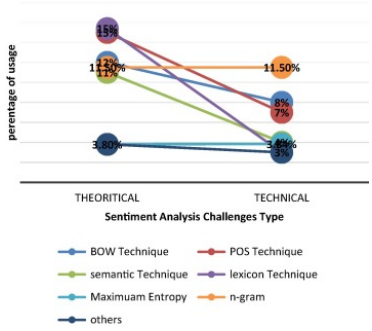


Figure 2.2: Sentiment Analysis Challenges Type and Techniques Type [[24]]

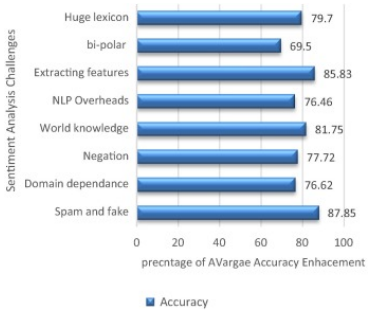


Figure 2.3: The improvement in accuracy results in sentiment analysis challenges[[24]]

us with sentiment analysis more accurately. Some of the renowned ones are LSTM, KNN, CNN, BERT, Multinomial Naive Bayes, Logistic regression, RNN and many more. The RNN stands for Recurrence Neural Network which is a prominent deep learning algorithm. It works following the working principle of neural networks. The concept of neural network was inspired from the structure of the human brain and how it works. Like the human brain, the neural network is designed based on the connections of neurons with one another which does the task of information processing [37]. Following the approach of neural networks RNN makes use of its internal memory to process information sequentially which allows it to execute a single task for all the elements of a sequence. The LSTM or long short term memory network is a special type of RNN model. It takes a chain type formation which contains four neural network layers. These networks together with memory blocks are called cells. Depending on the structure, LSTM cells can have multiple directions and dimensions [35]. The commonly used LSTM are unidirectional but there are bidirectional and multidirectional LSTMs as well where multiple cells are stacked creating a hierarchy.
The Convolutional Neural Network or CNN is another advanced algorithm mostly used in the field of image and speech recognition. However, its usage in natural language processing is on the rise as well. Being a multilayer network, the CNN algorithm gives the output of one layer as the input for the next layer [33]. The weights in the networks are assigned in accordance with the training data. To reduce the complexity in computation, it uses the polling layers to minimize the output size of one stack layer to the next, preserving only the important information .

2.1 Text analysis for human emotion detection

Feelings can be communicated verbally, physically, or in writing [28]. Expressions of feeling can also take the form of actions. The identification of material based on

the principles of deep learning presents a challenge when it comes to the recognition of emotions in text sources. Emotions play an essential role in the day-to-day functioning of humans [43]. Emotions are characterized by intuition that is distinct from rational cognition or grasp of the facts. The ability of an individual to reason about and control how they respond to incentives might be impacted by their emotional state [30].
One of the hallmarks of sophisticated conversation is the ability to articulate one's feelings [31]. Emotion detection technology can help find illegal motivations [22]. An individual's voice, looks, and written words can all be used to psychologically portray an emotion. Although progress has been achieved in emotion recognition using speech and facial expressions, a text-based emotion detection framework is still needed [36]. For data analysis purposes, it is extremely useful to be able to pinpoint where in a document human emotions are expressed [44]. Emotions such as joy, sorrow, anger, euphoria, hatred, fear, etc. are displayed.
Sometimes, this condition is associated with a change in the nature or the context of a person's conscious thought. Again, people's private views and the environment around them are determined by their emotional responses such as pleasure, grief, fear, anger, surprise, and so on [8]. Emotions are an essential component of a person's identity and their life experience. Words, short sentences, videos of facial expressions, lengthy messages, text, and emoticons are just some of the numerous forms of feedback that can be used to generate emotional responses. There are many other forms of feedback as well. These methods of data entry are not consistently supported by software [29].

2.1.1 Sentimental emojis and emoticons

Emojis give people the ability to communicate their feelings via the use of nonverbal cues and straightforward graphics by giving users the ability to send and receive emojis and by enabling users to send and receive emojis. Because of this, users are able to communicate with one another in a more efficient manner. It is likely that the use of emojis will make it simpler for us to understand the tone of a message that has been delivered to us in the form of a text or tweet, as well as to differentiate between content that communicates positive and negative views. This would be made possible by the fact that emojis are able to express a wide range of emotions and expressions. Emojis are used regularly on a variety of websites, including Twitter, which is among them [47].When sending out tweets, users of Twitter almost always use a significant number of emoticons that have never been seen before, and this tendency continues almost universally. The emotions that are communicated by these emoticons may be happy, sad, or neutral, depending on the context in which they are used. It was also able to determine the average amount of emojis that each user takes in on a daily basis by making use of a wide variety of statistical data. The system was successful in achieving this goal. A significant number of the messages mentioned a particular use of emoticons. People who have made at least one attempt are more likely to use fewer emoticons overall. In comparison to other sorts of writing, any emotion tends to contain a smaller number of emoticons. They compared the number of times emoticons occurred in texts that had thoughts to those that did not contain such ideas. Emojis that depict a person with folded hands, a face with no expression, a face that is pleading, a face

that is crying, a face that is yelling, a face that is bewildered, and a face that does not have a mouth are frequently used by people who are depressed. Emojis such as the winking face, the smiling face with heart eyes, the crying face, the laughing face, the red heart, the clapping hands, and the winking face are used by persons who are not depressed. According to these studies, individuals with depression use more negative emoticons. Happy people use more positive emoticons. Depressed people use more negative emoticons.

2.1.2 Extracting character traits

Every single person who is included in the database contains a one-of-a-kind combination of distinctive qualities that help to set them apart from the other individuals in the database [41]. These qualities serve to differentiate each individual from the others. It functions as a stand-in for the several facets of character that are associated with a certain attribute or trait of the individual. On the basis of this information, one is able to draw inferences as to whether the person is more of an extrovert or an introvert in their personality. When attempting to evaluate the personality traits of a user, one of the criteria that is taken into consideration is the polarity of the user's tweets, in addition to a few other pieces of statistical data. This is done in conjunction with a few other pieces of information. We take into consideration features of people's personalities such as how amiable they are, how optimistic they are, how socially active they are, how neurotic they are, and how much time they spend watching television. Before beginning to calculate the values of the various personality traits for each user, it is vital to first ascertain the polarity of the entire corpus of the set that is under suspicion of being suspicious. This should be done before beginning to compute the values of the various personality qualities. Before beginning to calculate the values of the different personality traits for each user, this step needs to be completed first. After this stage has been finished, it will finally be possible to compute the values of each user's unique mix of personality traits. This will be possible after step one has been finished. After then, the tone of each individual user's tweet was dissected and compared to the overarching tone of the corpus so that inferences could be drawn. On the other hand, if the majority of a person's tweets are critical, then it gives the impression that they have a neurotic outlook on life. It is believed that a person has an optimistic attitude on life if the majority of the tweets that they post are upbeat. In a manner that is equivalent to this, the overall polarity of the corpus has a lot of similarities with the polarity of a tweet that originated from a person who is normally agreeable. This is as a result of the fact that the corpus was generated by a group of people who, on the whole, held views that were comparable to one another. Even if the overall polarity of the corpus is neutral, a tweet that originates from someone who is both nice and engaged will have a positive polarity. This is the case even if the polarity of the corpus as a whole is neutral.

2.1.3 Social Searcher

A social media toolkit known as Social Searcher is offered for sale at a price that is seen as being within acceptable parameters. This social media toolkit provides a summary of both positive and negative sentiment by using the phrases and hashtags

that are contributed by users. It sorts the information in a similar manner like the raw collected social media contents and identifies the implicit keywords and hashtags to generate a basic summary.

2.1.4 Lexalytics

Lexalytics use natural language processing to detect a person's text or message and then it performs sentiment analysis to uncover that person's emotion [2]. It is an internet based sentiment analyzing software that extracts insights from texts that can describe a person's mood while communicating.

2.1.5 Extraction of features

Feature extraction is one of the sentiment analysis tools that is designed to find out the characteristics from a person's profile by a thorough checking. This is because it is the stage that actually extracts the features [15]. This is as a result of the fact that a full understanding of the user is required in order to obtain a high degree of accuracy in the identification of suicidal thoughts in a user. This is owing to the fact that a full understanding of the user is required in order to prevent the user from harming themselves. "Parsing" was performed on the tweets that were evaluated for this study so that a wide variety of data points could be extracted from within the tweets themselves. These characteristics included, but were not limited to, elements of social media, emoticons, emotion analysis, TF-IDF, statistics, topic-based characteristics, and temporal characteristics, amongst others. In addition, these features were not limited to any particular historical period at any point in history.

2.1.6 Social media features

This section contains the information that was typically taken from a user's profile. This information includes the user's name, the language that he used to communicate, the friend count, how his profile description, country, date the profile was created, time zone, profile photo, follower, and following, in addition to a number of other pieces of information that were typically taken from a user's profile. The polarity and subjectivity of a profile are two examples of the kinds of additional characteristics and classifications that may be produced by using these features and characteristics as building blocks. These are just two examples of the kinds of other characteristics and classifications that can be constructed [45].These are only two examples of the many other sorts of traits and classifications that can be built. It is likely that the polarity will place a high value on the user profile description. This is because the user profile description embodies the persona of the individual and will play a significant role in the process of classifying persons. This is due to the fact that the user profile description will serve as a representation of the individual's persona.

2.1.7 Tweet information

The statistical measurements provided in this feature area are the result of analyzing tweets posted by users and were gathered as a consequence of that analysis. For the

purpose of obtaining these measurements, this analysis was carried out [46]. Some of the messages have vocabulary that is easy to understand and get right to the point, but some of the communications contain terminology that is more difficult to understand and contain paragraphs that are considerably lengthier. Consideration is given to each and every facet of an individual user's Twitter activity. This contains the number of tweets per user that are related with suicide as well as the proportion of a user's total postings that are associated with either happiness or sorrow. It also includes the length of time that tweets are posted. According to the findings of the study, users who said that they had not given any consideration to the possibility of ending their own lives sent, on average, longer messages than users who did not indicate that they had given any consideration to the possibility of ending their own lives. This is because users who are contemplating are more likely to have issues with their mental health or social problems, both of which are represented in the postings that these users generate. This is because users who are contemplating are more likely to have issues with their mental health or social problems. Furthermore, those who are considering these options are more likely to have access to the internet. This is owing to the fact that individuals who are using substances and contemplating taking their own lives are more likely to have issues with their mental health as well as social issues.

2.1.8 Personality trait extraction

The dataset includes users, each and every one of whom have a personality that is distinct from the others [26]. It is a depiction of the many facets of a person's personality that are connected to a certain quality of that individual. Because of this knowledge, it is feasible to determine if the person in question is an introvert or an extrovert. A computation is performed, which, in addition to a variety of statistical criteria, takes into account the polarity of the user's tweets. The purpose of this calculation is to estimate the personality characteristics of each individual user. The following five primary personality qualities are discussed in this article: agreeableness, optimism, active sociality, neuroticism, and spectatorship. It is required to first determine the polarity of the entire corpus of the collection that is being suspected before one can proceed with the task of computing the values of the numerous personality traits that are associated with each individual user. The purpose of the current investigation is to learn about the features of social media users who have posted suicide or otherwise dismal thoughts or feelings in their postings. This study will be taken into consideration in order to accomplish this goal. The overarching purpose of the study is to investigate the connections between the various user attributes. According to the findings of an examination of the dataset, personality variables were found to be significantly associated with all different kinds of thoughts. The normative sample has a tendency to have users who are more agreeable and optimistic than the population at large. These findings lend even more credence to the theory that a person's personality traits are among the primary factors that contribute to their downfall in the long run. The polarity of each tweet was analyzed and compared to the corpus. If a user has a majority of positive tweets, it indicates that they are optimistic; otherwise, it indicates that they are neurotic. The polarity of the corpus aligns with the tweet of a user who is on board. When the corpus is in a state of neutrality, active and social people tweet

in a favorable manner.

2.2 Related works

There are two main camps within the field of emotion recognition analysis that have traditionally been used: One, using a lexicon, and two, using machine learning. To classify how one feels, lexicon-based approaches use already-compiled word lists [6]. Research projects are currently being conducted with the purpose of determining whether or not there is a distinction in the thoughts that are entertained by young people of either gender who have a propensity toward emotional tendencies [18].
The findings of the study also shed light on which gender is more likely to attempt to end their own life by taking their own life. The answer can be found in the previous sentence. This research was conducted using data collected from students in a variety of educational institutions. Anxiety, a sense of not belonging, and feeling alone are the three basic aspects that the writers identified as having the ability to have an effect on the mental health of students. The authors also acknowledged that these three primary elements have the capacity to have an influence. Isolation was cited by the writers as another critical problem in their work. Using machine learning, decision criteria for emotion recognition can be inferred from a labeled database of training examples [3], [5]. Researchers have attempted a variety of techniques to forecast how a reader will feel while reading a narrative text. Methods such as Naive Bayes (NB) [9], Random Forest (RF) [14], [19], Support Vector Machine (SVM) [24–26], [17], and Logistic Regression (LR) [7] all utilize the adaptability of machine learning. All of which are commonplace in written communication. Rarely, but occasionally, these classifiers are restricted to affect signals from emotion lexicons [11]. In addition to this, they provided pertinent coping skills that could be applied in the case of a crisis that was of a similar nature to the one that they had encountered. These ways could be utilized in the event that a crisis was of a comparable character to the one that they had experienced. An article titled "11" that was produced by the writers provides a summary of the numerous and varied types of suicidal behavior that may be found in the world today. The authors undertook study on the epidemic, looking at the elements that increase the risk of suicide behavior as well as the characteristics that decrease the risk of suicidal behavior. According to the findings of a different study, children are exposed to a variety of different risk factors, any one of which could result in the development of this condition [16]. At this time, the act of suicide is one of the most significant challenges that humanity must contend with in our modern day. Many statistics relating to suicide, suicidal ideation, and attempted suicide are provided by the authors of the article , which may be found here [21]. The writers of , which is another excellent book on the subject, made use of NLP in their research in order to address suicidal ideation as well as attempts at suicide [2]. is one of the best works on the topic . This is the location of [23].
The authors of the study referenced in came up with a method for estimating the suicide risk of users of social media platforms who speak Spanish [42]. They did this by collecting data that was observable, logical, and multimodal from a number of different social platforms and developing algorithms that could identify users who were going for suicidal intention and these types of thoughts. In order to accomplish this, they collected data from the various social sites, and the authors of developed a

method for calculating the risk of suicide posed by users who communicate primarily in Spanish on various social media platforms [13].
In order to achieve this goal, data was gathered from a variety of social networking sites, and then it was examined in an effort to identify individuals who were having thoughts of ending their own lives by killing their own lives. employed both human coders and a machine learning classifier to assess whether or not the amount of concern that was indicated in a tweet about suicide could be discovered only by the text of the tweet itself [20]. The study was conducted to determine whether or not a tweet about suicide could be used to detect the level of concern that was conveyed in the tweet. This test's objective was to determine whether or not it was possible, based solely on the tone of language employed in a tweet, to deduce the degree of worry that the tweeter was attempting to convey.
Using Twitter users' tweets, the authors of [32] mined six feature groups constituted of depression-related criteria from clinical and online social behaviors to create a multimodal depression dictionary learning model that may be used to identify depressed users of Twitter. Users of Twitter who are depressed can be identified with this model. Twitter users experiencing depression can be identified using this technique. Using a multimodal depression dictionary learning model, this approach has the potential to be utilized to diagnose Twitter users going through depressive episodes.The authors of [48] argue that anxiety and depression are linked to restlessness, a difficulty to sleep, and aberrant thought processes. They developed a model for identifying anticipatory melancholy that was based not just on linguistic indicators but also on real-time tweets and posting patterns. This technique was able to detect users who were experiencing anxiousness and despair and determine which individuals were going through these feelings.

Chapter 3

Fundamentals of Work

3.1 Dataset

3.1.1 Data Collection

Our work begins with the collection of some datasets as well as possible. Dataset plays an important role in the research because the prediction accuracy of the target variable quite depends on how refined the dataset is. As per our objective, we collected datasets from Kaggle, that comprise sentences with six basic human emotions. One of the datasets is made of tweet emotions for more refined data to express human sentiment. To be precise about data validation, the datasets were separated into train, test and val parts.

3.1.2 Dataset Analysis

From the amount of data, it can be visualized that people mostly tend to share their happy or joyful moments (about 32%) on social platforms. Sadness comes in second position (about 29%), implying that there happens quite a lot of sorrowful events in people's lives which they want to express. Followed by these two, there are also anger, fear, love and surprise elements in their conversation or speech less in proportion as the data suggests. We have worked with 3 different datasets to compare between them about which model gives better accuracy. Two of our datasets contained around 20,000 data and the other one had about 40,000 data. The data of the first two datasets were not collected pragmatically, rather arbitrarily as they had only two columns- text and emotion. There were 6 different sentiments in them- happy, sadness, love, anger, surprise and fear. The third dataset comprised tweet emotions and it had three columns named in- tweet-id, sentiment and content. We believed this dataset would be more reliable in case of sentiment analysis. Additionally, it had 13 unique sentiments as in happiness, sadness, love, anger, surprise, fear, worry, fun, relief, hate, empty, enthusiasm, boredom, and neutral.

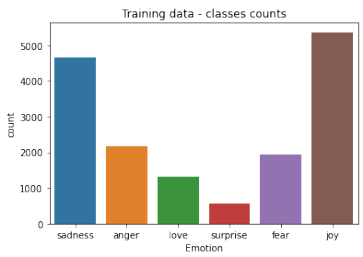


Figure 3.1: Train

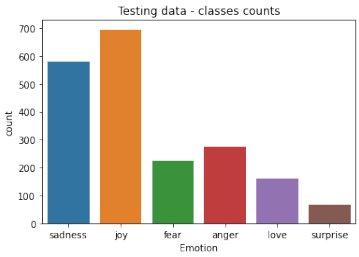


Figure 3.2: Test

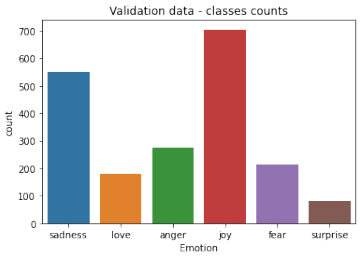


Figure 3.3: Validation

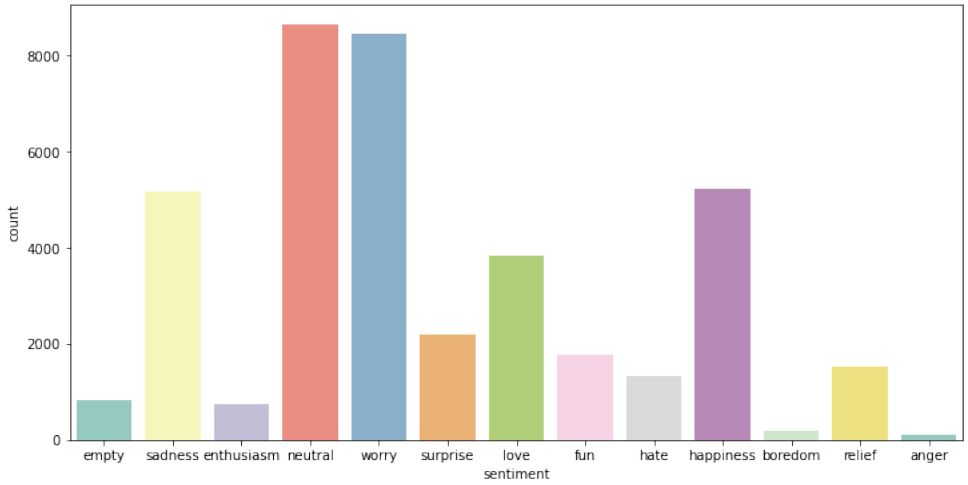


Figure 3.4: Sentiment Analysis

3.2 Proposed Methodology

3.2.1 CNN-LSTM

CNN-LSTM is a popular deep-learning algorithm that combines the features of CNN and LSTM. Both CNN and LSTM are individually great algorithms in the fields of text classification, speech recognition and video analysis but we thought of using a hybrid algorithm to see how it produces outcome, compared to the individual algorithm. In this method, CNN extracts patterns from input data and gives that pattern as input to the LSTM network for sequence modeling and predictions []. The LSTM contains hidden layers which it updates taking input from CNN. LSTM gates (input, forget, output) control information flow and update hidden state [34] Mathematically,

$$h_t = LSTM(c_t, h_{t-1}) \tag{3.1}$$

h_t is the LSTM's hidden state at time step t, c_t is the CNN's output, and h_{t-1} is the prior hidden state.

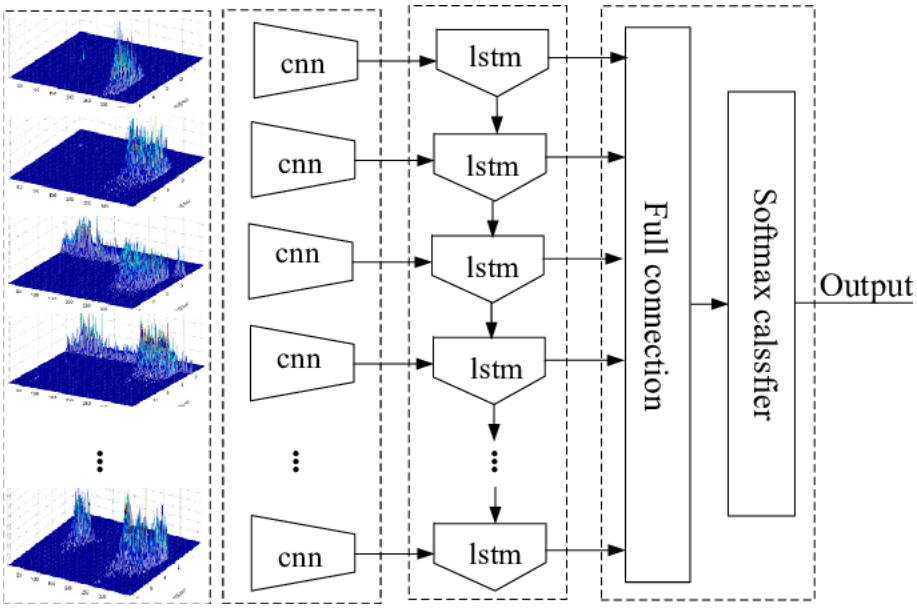


Figure 3.5: CNN-LSTM Architecture [[39]]

3.2.2 LSTM

LSTM is another deep learning algorithm which is basically a type of RNN. The LSTM network consists of memory cells, gates, and next time step. Firstly, the model takes text input as vector, then the input is processed through three gates. The forget gate takes the input and determines whether the value from the previous step should be stored in the memory cell or removed. The Input gate applies sigmoid activation function to the current data and stores it in the memory cell. Next, the memory cell is updated with the information from the forget gate and input gate. Followed by, the output gate takes the sequence of the updated memory cell and feeds it as the input of the next time step. After all the time steps, the prediction is generated [38] Mathematically:

$$i_t = sigmoid * (W_i * x_t + U_i * h_{t-1} + b_i) \tag{3.2}$$

$$C_t = i_t * tanh * (W_c * x_t + U_c * h_{t-1} + b_c) \tag{3.3}$$

Cell state, input gate, and bias weights.

f_t decides how much of C_{t-1} is kept.

Mathematically:

$$f_t = sigmoid(W_f * x_t + U_f * h_{t-1} + b_f) \tag{3.4}$$

$$C_t = C_t + f_t * C_{t-1} \tag{3.5}$$

The output gate, o_t, controls how much cell state, C_t, information is used to generate the hidden state, h_t. Mathematically:

$$o_t = sigmoid(W_o * x_t + U_o * h_{t-1} + b_o) \tag{3.6}$$

$$h_t = o_t * tanh(C_t) \tag{3.7}$$

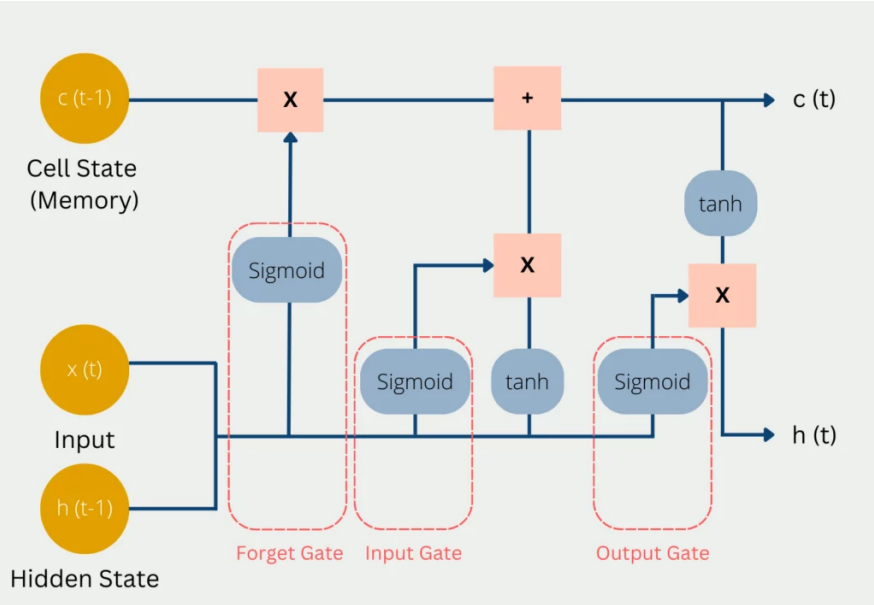


Figure 3.6: LSTM Architecture[[49]]

3.2.3 RNN

RNN or Recurrent Neural Network is one of the most effective algorithms when it comes to natural language processing. It has a hidden state that is updated at each time step. Taking an input, the RNN combines it with the previous hidden state and applies an activate function to generate a new hidden state. It continues until all elements of the input sequence are iterated.

If x_t is the input, hidden state is h at time step t, then:

$$h_t = f * (W * x_t + U * h_{t-1}) \tag{3.8}$$

where f is the sigmoid function and W is the weight. Using V weights, the new hidden state, h_t , produces an output, y_t :

$$y_t = g * (V * h_t), \tag{3.9}$$

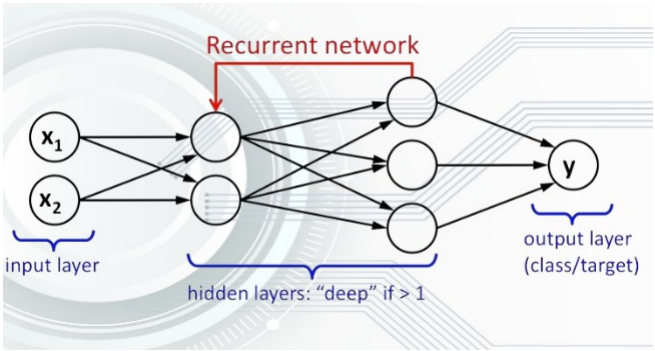


Figure 3.7: RNN Architecture

3.3 Work Flow

As per our objective, at first we collected data regarding people's sentiment. Our dataset contained texts with innate sentiment. Next, we cleaned the dataset as there remained many unnecessary characters in the raw data. It reduced the data noise and improved the quality of input. In the preprocessing part, we encoded, tokenized and embedded data to prepare it for machine learning algorithms. After data was prepared, we inputted it in three different algorithms which are, CNN-LSTM, LSTM and RNN. In addition, we compared the accuracy of the three models to find out which model comes in more profound in detecting human sentiment.

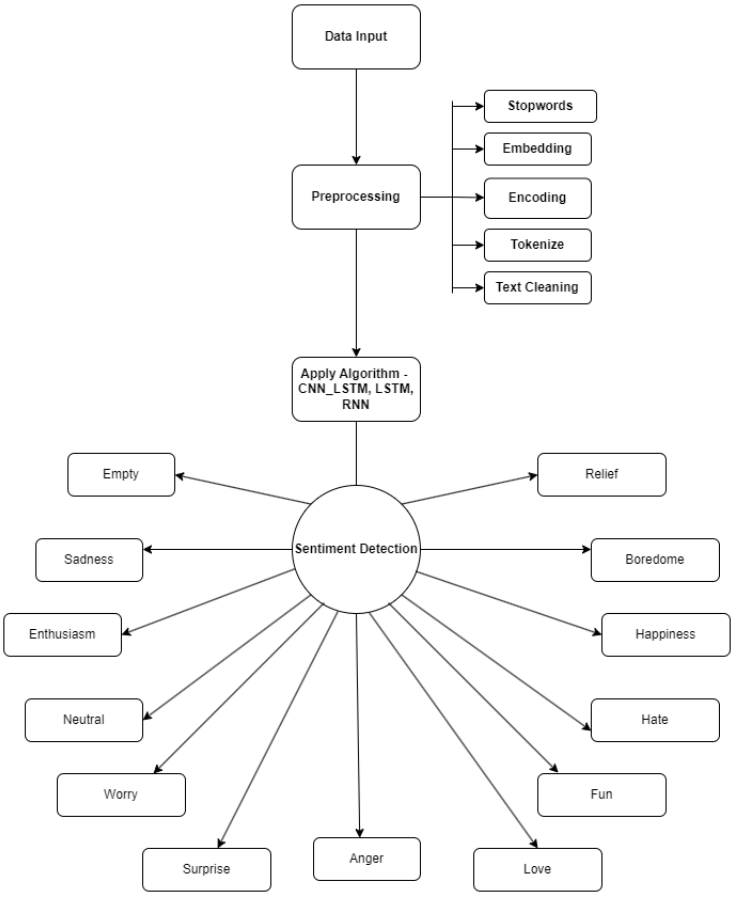


Figure 3.8: Work Flow Diagram

Chapter 4

Model Implementation

We've developed a Convolutional Neural Network (CNN) integrated with Long Short Term Memory (LSTM), Recurrent Neural Network (RNN), and Long Short Term Memory (LSTM) models that can identify various sentiment conditions from text input. In order to determine which text data represent happiness, sadness, fear, and anger and which do not, this classification issue will employ features to find distinct patterns and sequences in the input text. This research methodology is compatible with classifier algorithms as our target consists of six different emotions.

4.1 Data Preprocessing

In order to preprocess data, we applied the following techniques:

4.1.1 Encoding

Encoding means converting categorical or non-numerical data into numerics for algorithms to process. There are various types of encoding that are used in machine learning algorithms. For our purpose, we used label encoding to transform our train, test and validation data into numerical values.

4.1.2 Tokenization

Tokenization is the process of splitting input texts into smaller individual units. These units are called tokens and they can be characters or words. They help the algorithms to detect the sequence, extract its features and analysis more accurately.
Dataset 1: Vocabulary size 14980
Dataset 2: Vocabulary size 19260
Dataset 3: Vocabulary size 45559

4.1.3 Embedding

Embedding means representing input texts in low-dimensional vector space. It is applied to extract semantic relationships between words in a sentence or document. We kept embedding size 128 as in 128 dimensions. Also, we defined max-length to be 100 as the maximum input size for the algorithms.

4.1.4 Stop Words

Stopwords is another preprocessing technique that is used to remove unnecessary words. It reduces data noise and enables the algorithm to perform better. We imported the stopwords module from the NLTK library for the English language. Here are some reasons why stop words are often removed in machine learning:

- Noise reduction
- Memory and Storage Efficiency
- Improved Computational Efficiency
- Improve feature Representation

4.1.5 Text Cleaning

When it comes to machine learning, natural language processing (NLP) activities, in particular, text cleaning (also called text preparation or text normalisation) is crucial. Machine learning models can be hindered by the presence of noise, irrelevant information, and inconsistencies in text data. To fix these problems and make the data more trustworthy, text cleaning methods are used. Here are some reasons why text cleaning is used in machine learning:

- Noise reduction
- Consistency and standardization
- Tokenization
- Stop word removal
- Lemmatization and stemming
- Handling spelling errors and abbreviations
- Removing rare or infrequent words

4.2 LSTM Model Implementation

The Keras Sequential API is used to build the model.

- In the first layer, known as the embedding layer, vocabulary that has been integer-encoded is transformed into dense vectors of a set size. It requires three parameters: input-length, embedding-dim, and vocabulary-size. Input-length denotes the length of input sequences, embedding-dim denotes the dimensionality of the dense embedding, and vocab-size denotes the size of the vocabulary (the total number of unique words).The LSTM layer with 128 units and return-sequences = "True" is the following layer. The configuration of this LSTM layer is set to return sequences as opposed to a single output. When stacking many LSTM layers, this is helpful.

- After the initial LSTM layer, a dropout layer with a dropout rate of 0.2 is added. Dropout is a regularization approach that helps avoid overfitting by randomly setting a portion of input units to 0 during training.

- There is an additional 64-unit LSTM layer added. Only the final output in the series will be returned because this layer does not have return-sequences = "True".

- The second LSTM layer is followed by a dropout layer with a dropout rate of 0.2.

- The addition of a dense layer with 128 units and a ReLU activation function. ReLU (Rectified Linear Unit) is a well-liked activation function that incorporates non-linearity, and dense layers are fully connected layers.

- After the dense layer, a dropout layer with a dropout rate of 0.2 is added.There is an additional dense layer with 64 units and a ReLU activation mechanism.There is an additional dropout layer added with a dropout rate of 0.2.

- 13 units of a dense layer are added. There is no defined activation function for this layer.

- The softmax activation function is added to an activation layer. As Softmax transforms the model's raw outputs into probability scores, it is frequently employed for multi-class classification issues.This model's overall architecture comprises of an embedding layer, two LSTM layers, and then a dropout layer for each layer. A dense layer with 13 units is then followed by a softmax activation layer for multi-class classification, followed by two dense layers with dropout layers in between.

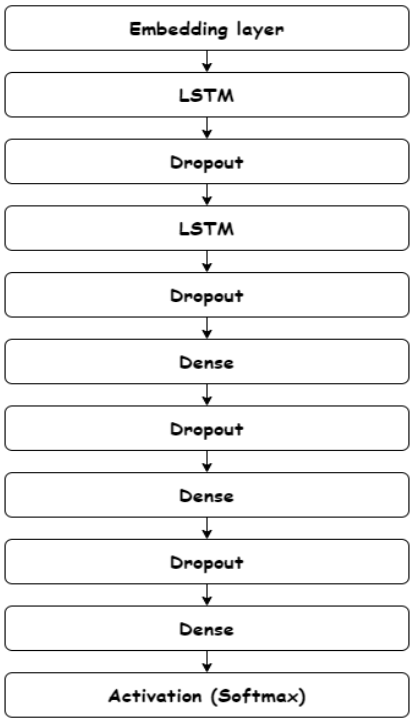


Figure 4.1: LSTM Model Layering

4.3 CNN-LSTM Model Implementation

The model is created using the Sequential API from Keras.

- The first layer is an Embedding layer, which is used to convert integer-encoded vocabulary into dense vectors of fixed size. It takes three parameters: vocab-size, embedding-dim, and input-length. The vocab-size represents the size of the vocabulary (the total number of unique words), embedding-dim represents the dimensionality of the dense embedding, and input-length represents the length of input sequences.

- The next layer is a 1D Convolutional layer (Conv1D) with 256 filters, a kernel size of 5, and 'valid' padding. It uses the ReLU activation function and has a stride of 1. Convolutional layers are commonly used for extracting local features from the input data.

- A MaxPooling1D layer with a pool size of 2 is added. Max pooling reduces the spatial dimensions of the input, preserving the most important features.

- BatchNormalization layer is added to normalize the activations of the previous layer.

- A Dropout layer with a dropout rate of 0.3 is added to prevent overfitting by randomly setting a fraction of input units to 0 during training.

- Another Conv1D layer with 128 filters, a kernel size of 5, and 'valid' padding is added. It also uses the ReLU(Rectified Linear Unit) activation function and has a stride of 1.

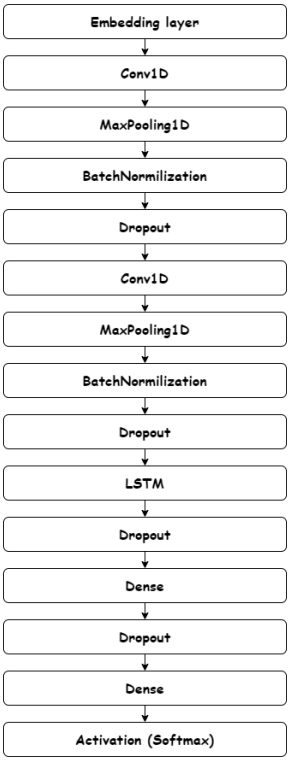


Figure 4.2: CNN-LSTM Model Layering

- Another MaxPooling1D layer with a pool size of 2 is added.
- Another BatchNormalization layer is added.
- Another Dropout layer with a dropout rate of 0.3 is added.
- An LSTM layer with 128 units is added. LSTM layers are a type of recurrent neural network (RNN) layer that can capture long-term dependencies in sequential data.
- A Dense layer with 32 units is added.
- A Dense layer with 13 units is added. This layer does not have an activation function specified.
- An Activation layer with the softmax activation function is added. Softmax is commonly used for multi-class classification problems as it converts the model's raw outputs into probability scores.

Overall, this model architecture combines both convolutional and recurrent layers. The convolutional layers are responsible for extracting local features from the input data, while the LSTM layer captures the temporal dependencies. The model also includes dropout and batch normalization layers to prevent overfitting and improve generalization. Finally, the model ends with a softmax activation layer for multi-class classification.

4.4 RNN Model Implemention

The model is created using the Sequential API from Keras.

- The first layer is an Embedding layer, which is used to convert integer-encoded vocabulary into dense vectors of fixed size. It takes three parameters: vocab-size, embedding-dim, and input-length. The vocab-size represents the size of the vocabulary (the total number of unique words), embedding-dim represents the dimensionality of the dense embedding, and input-length represents the length of input sequences.
- The next layer is a Simple RNN layer with 128 units and ReLU activation function. SimpleRNN is a type of recurrent neural network (RNN) layer that processes sequential data by maintaining a hidden state and updating it at each time step.
- A Dropout layer with a dropout rate of 0.3 is added to prevent overfitting by randomly setting a fraction of input units to 0 during training.
- A Dense layer with 64 units and ReLU activation function is added. Dense layers are fully connected layers, and ReLU is a popular activation function that introduces non-linearity
- Another Dropout layer with a dropout rate of 0.3 is added.
- Another Dense layer with 32 units and ReLU activation function is added.
- Another Dropout layer with a dropout rate of 0.3 is added.
- Another Dense layer with 16 units and ReLU activation function is added.
- Another Dropout layer with a dropout rate of 0.3 is added.
- A Dense layer with 13 units is added. This layer does not have an activation function specified.
- An Activation layer with the softmax activation function is added. Softmax is commonly used for multi-class classification problems as it converts the model's raw outputs into probability scores.

Overall, this model architecture consists of an embedding layer followed by a Simple-RNN layer. After that, there are several dense layers with dropout layers in between, and finally, a dense layer with 13 units followed by a softmax activation layer for multi-class classification.

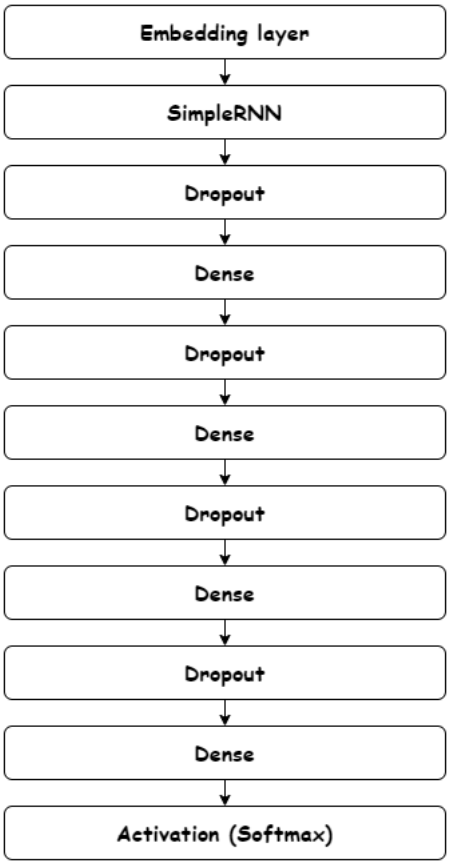


Figure 4.3: RNN Model Layering

Trainable Parameters of Different Models for Each Dataset			
Datasets / Models	Dataset - 01	Dataset - 02	Dataset - 03
CNN-LSTM	481,613	2,536,422	1,104,742
LSTM	2,269,638	837,958	301,069
RNN	683,862	2,115,542	189,645

Figure 4.4: Total Trainable Parameters

Chapter 5

Result Analysis

5.1 Dataset-1

5.1.1 CNN-LSTM

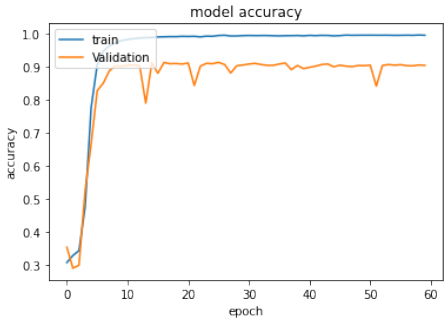


Figure 5.1: Accuracy

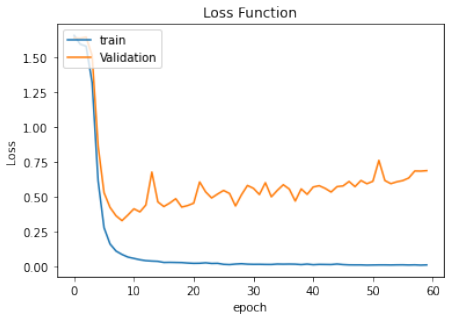


Figure 5.2: Loss-Function

This combined architecture yielded highly accurate sentiment detection, with an achieved accuracy of 0.905 which is equivalent to 92.90%. Furthermore, the data loss for this method was measured at 0.688 . Also, we can see a little fluctuation in accuracy and validation function which indicates there has been complexity while prediction.

5.1.2 LSTM

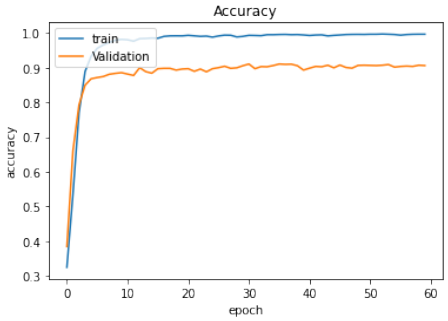


Figure 5.3: Accuracy

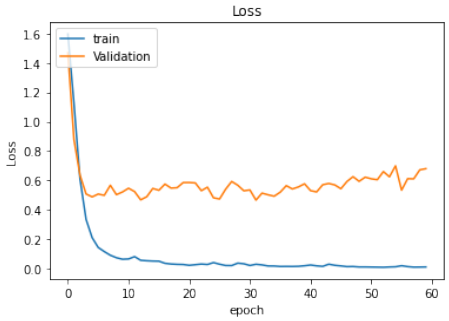


Figure 5.4: Loss-Function

The LSTM model performed with almost the same accuracy which is 0.906 and the same loss, that is 0.679. But the graph is quite flat out . That means the model was built and trained quite properly and its parameters functioned very well in predicting the sentiment.

5.1.3 RNN

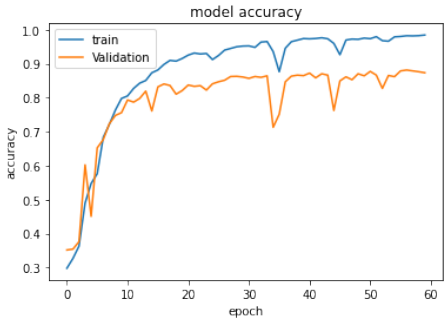


Figure 5.5: Accuracy

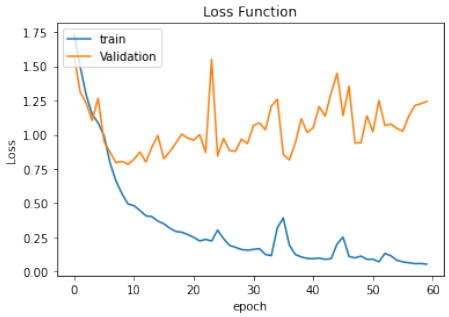


Figure 5.6: Loss-Function

The RNN model didn't give above 90% accuracy like the other two models. It produced an accuracy of 0.87 and loss function of 1.24. Moreover, there are a number of spikes in the graph, suggesting that this model could not perform well in terms of both accuracy and loss. The reason might be some discrepancies in training or the model was underfitted. Hence, it resulted in unpredictability.

5.2 Dataset-2

The accuracy and loss graphs for the CNN-LSTM, LSTM and RNN models using the emotion intext dataset are displayed below for various time periods.

5.2.1 CNN-LSTM

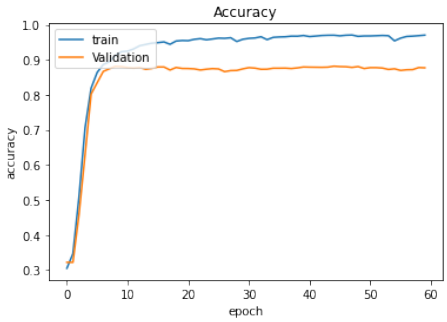


Figure 5.7: Accuracy

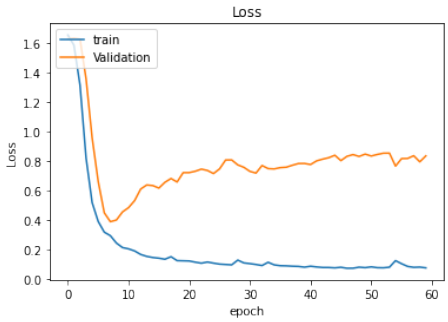


Figure 5.8: Loss-Function

The CNN-LSTM model was able to carry out human sentiment detection with an accuracy of 0.87 while the loss function here was 0.83 . The graph is very linear, suggesting that there wasn't much complexity and the model gave a decent performance.

5.2.2 LSTM

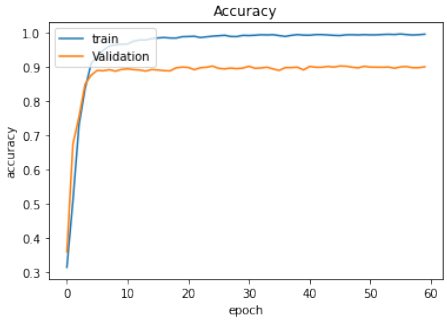


Figure 5.9: Accuracy

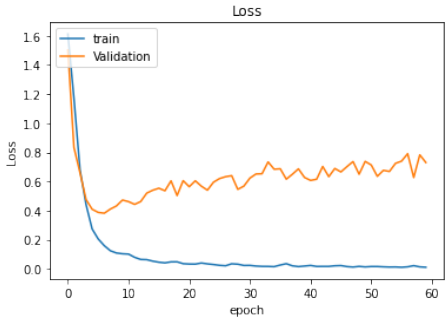


Figure 5.10: Loss-Function

The LSTM model produced output with accuracy of 0.9 with a loss function of 0.73. Though, there seems to be a little fluctuation in the graph but the overall result is up to the mark.

5.2.3 RNN

The RNN model carried out the prediction quite well with an accuracy of 0.90 but the high loss function which is 2.39 makes overall the performance unsatisfactory because it has lost too much data. Also, the spikes in the loss function graph shows an unstable performance.

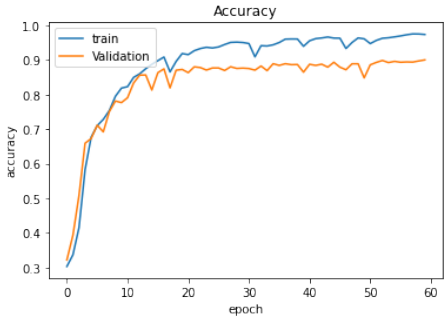


Figure 5.11: Accuracy

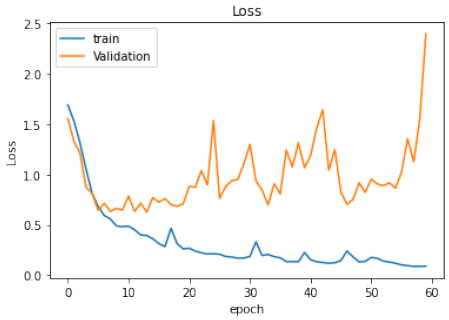


Figure 5.12: Loss-Function

5.3 Dataset-3

5.3.1 CNN-LSTM

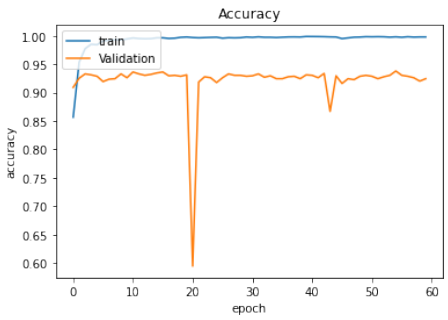


Figure 5.13: Accuracy

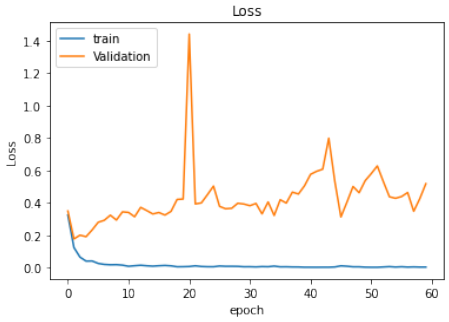


Figure 5.14: Loss-Function

The LSTM model performed excellently with an accuracy of 0.92 and loss function of 0.51. Though there is a sudden spike in the graph as the accuracy declined at a great measure but also rose immediately. The loss function seems a little fluctuating but extensively, the model gave a stable performance.

5.3.2 LSTM

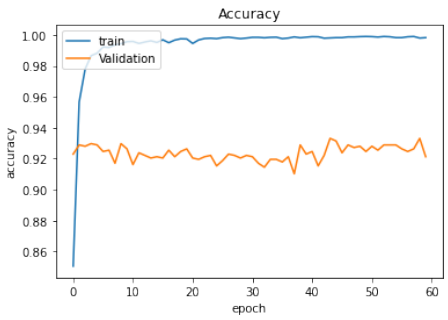


Figure 5.15: Accuracy

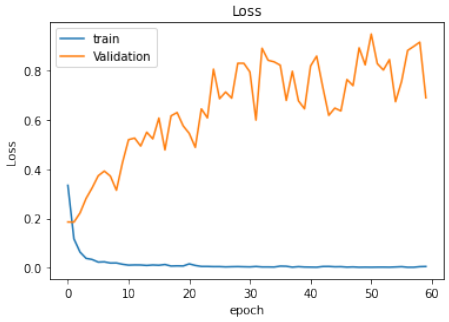


Figure 5.16: Loss-Function

The LSTM model also yielded high accurate output with 0.92 accuracy and 0.82 data loss. Similar to the hybrid model, here appears to be a little bit fluctuation in the accuracy graph but the loss function has too much oscillation.

5.3.3 RNN

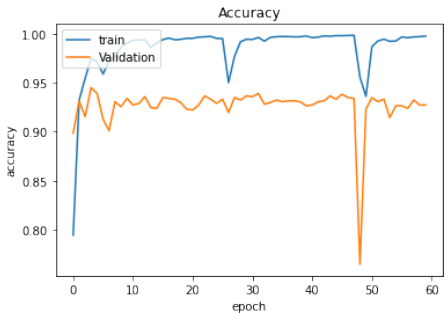


Figure 5.17: Accuracy

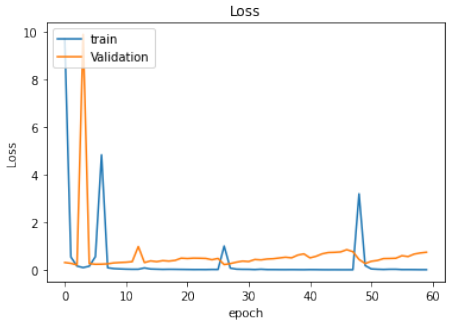


Figure 5.18: Loss-Function

The RNN model delivered good accuracy and loss function this time as compared to it's previous performance on other datasets. It brought about an accuracy of 0.92 and loss function of 0.74. There are few fluctuations and one sudden spike in both of the graphs which probably is because of inappropriate parameter setting that leads to the model's inability to capture input sequences and patterns.

Accuracy and Loss-function table for Dataset-1:

Model	Validation-Accuracy	Validation-Loss
CNN-LSTM	0.9050	0.6885
LSTM	0.9060	0.6795
RNN	0.8735	1.2422

Accuracy and Loss-function table for Dataset-2:

Model	Validation-Accuracy	Validation-Loss
CNN-LSTM	0.8772	0.83772
LSTM	0.9005	0.7304
RNN	0.9003	2.3941

Accuracy and Loss-function table for Dataset-3:

Model	Validation-Accuracy	Validation-Loss
CNN-LSTM	0.9246	0.5178
LSTM	0.9213	0.6903
RNN	0.9272	0.7419

5.4 Model Comparison

As we found out after result, for the third dataset, all the models gave better output in terms of accuracy and loss. One possible reason might be that this dataset is more refined than the other two. If we consider the overall performances of all three models on all datasets, it turns out that LSTM outperformed them all. It not only gave over 90% accuracy but also minimum loss function. The graphs of the LSTM model are also the most linear and have least fluctuations. So, it can be said that all parameters and functions of LSTM could fit the values of all different datasets properly and trained very well.
After LSTM, our hybrid model CNN-LSTM gave a decent performance. Other than dataset 2, this model delivered above 90% accuracy. The loss functions were around 0.5 - 0.6. The graphs seemed to have fluctuated little.
Lastly, the RNN model didn't perform well as its loss function is too high, meaning the model lost too much data in training. Though the accuracy was decent- around 90%, too many fluctuations and spikes in the graphs show the instability of the model.
So, considering comprehensive performances of all three models, we would like to recommend the LSTM model as the most useful means to sentiment analysis.

$$LSTM > CNN - LSTM > RNN$$

Chapter 6

Conclusion

Sentiment Analysis is a major research area in natural language processing. Using deep learning, human emotion is detected in various ways. Our research pursued the method to identify from text input. We took twitter data to work on as a refined source of social media content. We applied one hybrid model CNN-LSTM and two other ones LSTM, RNN to detect sentiments from different arrangements of data with a view to developing their precision. If the progress keeps going, this research could be a helping hand for developing the algorithms further to create a system to detect emotion more precisely. Where sentiment analysis is needed for social purposes and betterment of people, these systems and researches will hopefully come to help.

6.1 Challenges

Apart from the model training and testing, there have been a few limitations and challenges we had to go through for the research. Firstly, the quality of the first two datasets is doubtful as they were compiled arbitrarily. Since the quality of the dataset is significant for better accuracy and outcome. Also their small size made them unreliable as the model needs enough data to extract linguistic patterns. Choosing the correct datasets from a wide range of available datasets was difficult. Secondly, training the model properly was a tough thing for us. Importing the correct libraries to train and evaluate the models for all three datasets was a challenge because different datasets require different dependencies and modules to train.
Inconsistency in data labeling is a common problem in natural language processing. One situation or sentence may not be deemed in the same ways by everyone. So, putting a sentiment label on a data or sentence will not be universal. Therefore, the models were trained only based on the collected data and so they might not perform the same if applied on a much broader aspect.
Linguistic analysis would be a big challenge for us. We have researched sentiment analysis only for the English language. But how the models may deliver outcomes from Bengali or some other language is ambiguous because every language has its own linguistic characteristics and patterns. So, the model's ability to perform on other languages is uncertain.

6.2 Future Work

We hope to stick with our work for further improvement and thereafter our objective is to acquire a more in-depth knowledge in the field of natural language processing (NLP) and sentiment analysis (SA) are undergoing significant expansion and to evaluate how relevant our study is to the progression of modern computer science. If our efforts were acknowledged and supported by teachers and mentors, we could continue our research for a longer period of time. We also have an aim to identify one special type of sentiment alongside the previous ones and that is suicidal. As it's a major issue in the context of today's society, our target is to develop such a model that can detect such emotion very precisely. We look forward to working on the researched models and developing them to an extent of even greater accuracy. Additionally, we plan to implement our models on bengali data as it's our mother tongue. Additionally, we plan to implement our models on bengali data as it's our mother tongue. There are not enough research works on bengali sentiment analysis and so our target is to extend our research implementations up to bengali.

Bibliography

[1] A. J. Joshi, "Natural language processing," *Science*, vol. 253, no. 5025, pp. 1242–1249, 1991, ISSN: 00368075, 10959203. [Online]. Available: http://www.jstor.org/stable/2879169 (visited on 05/21/2023).

[2] W. Bunney, A. Kleinman, T. Pellmar, S. Goldsmith, *et al.*, "Reducing suicide: A national imperative," 2002.

[3] T. Danisman and A. Alpkocak, "Feeler: Emotion classification of text using vector space model," in *AISB 2008 convention communication, interaction and social intelligence*, vol. 1, 2008, pp. 53–59.

[4] I. S. Rubinstein, R. D. Lee, and P. M. Schwartz, "Data mining and internet profiling: Emerging regulatory and technological approaches," *U. Chi. L. Rev.*, vol. 75, p. 261, 2008.

[5] S. Chaffar and D. Inkpen, "Using a heterogeneous dataset for emotion analysis in text," in *Advances in Artificial Intelligence: 24th Canadian Conference on Artificial Intelligence, Canadian AI 2011, St. John's, Canada, May 25-27, 2011. Proceedings 24*, Springer, 2011, pp. 62–67.

[6] S. M. Mohammad, "From once upon a time to happily ever after: Tracking emotions in mail and books," *Decision Support Systems*, vol. 53, no. 4, pp. 730–741, 2012.

[7] M. Purver and S. Battersby, "Experimenting with distant supervision for emotion classification," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, pp. 482–491.

[8] R. Alazab, "0038 children below 5 years of employed mothers are less exposed to acute poisoning in alexandria, egypt," *Occupational and environmental medicine*, vol. 71, no. Suppl 1, A63–A63, 2014.

[9] M. Hasan, E. Rundensteiner, and E. Agu, "Emotex: Detecting emotions in twitter messages," 2014.

[10] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams engineering journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[11] S. Wen and X. Wan, "Emotion classification in microblog texts using class sequential rules," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 28, 2014.

[12] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, 2015.

[13] B. O'dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, and H. Christensen, "Detecting suicidality on twitter," *Internet Interventions*, vol. 2, no. 2, pp. 183–188, 2015.

[14] D. Gordeev, "Detecting state of aggression in sentences using cnn," in *Speech and Computer: 18th International Conference, SPECOM 2016, Budapest, Hungary, August 23-27, 2016, Proceedings 18*, Springer, 2016, pp. 240–245.

[15] A. P. Jain and P. Dandannavar, "Application of machine learning techniques to sentiment analysis," in *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, 2016, pp. 628–632. DOI: 10.1109/ICATCCT.2016.7912076.

[16] A. Bandhakavi, N. Wiratunga, D. Padmanabhan, and S. Massie, "Lexicon based feature extraction for emotion text classification," *Pattern recognition letters*, vol. 93, pp. 133–142, 2017.

[17] D. Chatzakou, A. Vakali, and K. Kafetsios, "Detecting variation of emotions in online activities," *Expert Systems with Applications*, vol. 89, pp. 318–332, 2017.

[18] P. Dholariya, *Suicidal tendency in youth in relation to their gender*. GRIN Verlag, 2017.

[19] A. Seyeditabari, S. Levens, C. D. Maestas, *et al.*, "Cross corpus emotion classification using survey data," *This paper was presented at AISB*, 2017.

[20] G. Shen, J. Jia, L. Nie, *et al.*, "Depression detection via harvesting social media: A multimodal dictionary learning solution.," in *IJCAI*, 2017, pp. 3838–3844.

[21] J. Bilsen, "Suicide and youth: Risk factors," *Frontiers in psychiatry*, p. 540, 2018.

[22] T. Chen, S. Ju, X. Yuan, *et al.*, "Emotion recognition using empirical mode decomposition and approximation entropy," *Computers & Electrical Engineering*, vol. 72, pp. 383–392, 2018.

[23] A. C. Fernandes, R. Dutta, S. Velupillai, J. Sanyal, R. Stewart, and D. Chandran, "Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing," *Scientific reports*, vol. 8, no. 1, pp. 1–10, 2018.

[24] D. M. E.-D. M. Hussein, "A survey on sentiment analysis challenges," *Journal of King Saud University-Engineering Sciences*, vol. 30, no. 4, pp. 330–338, 2018.

[25] A. Richterich, "Big data: Ethical debates," in *Big Data Agenda: Data Ethics and Critical Data Studies, The*, University of Westminster Press London, 2018, pp. 33–51.

[26] A. Shelar and C.-Y. Huang, "Sentiment analysis of twitter data," in *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, IEEE, 2018, pp. 1301–1302.

[27] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, e1253, 2018.

[28] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, and P. Agrawal, "Understanding emotions in text using deep learning and big data," *Computers in Human Behavior*, vol. 93, pp. 309–317, 2019.

[29] S. Dash, A. K. Luhach, N. Chilamkurti, S. Baek, Y. Nam, *et al.*, "A neuro-fuzzy approach for user behaviour classification and prediction," *Journal of Cloud Computing*, vol. 8, no. 1, pp. 1–15, 2019.

[30] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," *Information Fusion*, vol. 49, pp. 69–78, 2019.

[31] E. Kanjo, E. M. Younis, and C. S. Ang, "Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection," *Information Fusion*, vol. 49, pp. 46–56, 2019.

[32] A. Kumar, A. Sharma, and A. Arora, "Anxious depression prediction in real-time social data," *arXiv preprint arXiv:1903.10222*, 2019.

[33] A. U. Rehman, A. K. Malik, B. Raza, and W. Ali, "A hybrid cnn-lstm model for improving accuracy of movie reviews sentiment analysis," *Multimedia Tools and Applications*, vol. 78, no. 18, pp. 26 597–26 613, 2019.

[34] A. U. Rehman, A. K. Malik, B. Raza, and W. Ali, "A hybrid cnn-lstm model for improving accuracy of movie reviews sentiment analysis," *Multimedia Tools and Applications*, vol. 78, pp. 26 597–26 613, 2019.

[35] M. Rhanoui, M. Mikram, S. Yousfi, and S. Barzali, "A cnn-bilstm model for document-level sentiment analysis," *Machine Learning and Knowledge Extraction*, vol. 1, no. 3, pp. 832–847, 2019.

[36] K. Shrivastava, S. Kumar, and D. K. Jain, "An effective approach for emotion detection in multimedia text data using sequence based convolutional neural network," *Multimedia tools and applications*, vol. 78, pp. 29 607–29 639, 2019.

[37] K. Smagulova and A. P. James, "A survey on lstm memristive neural network architectures and applications," *The European Physical Journal Special Topics*, vol. 228, no. 10, pp. 2313–2324, 2019.

[38] R. C. Staudemeyer and E. R. Morris, "Understanding lstm–a tutorial into long short-term memory recurrent neural networks," *arXiv preprint arXiv:1909.09586*, 2019.

[39] L. Wu, J. Ding, Z. Li, Y. He, and Y. Zhang, "Research on transformer partial discharge uhf pattern recognition based on cnn-lstm," *Energies*, vol. 13, p. 61, Dec. 2019. DOI: 10.3390/en13010061.

[40] M. Marong, "Sentiment analysis in e-commerce: A review on the techniques and algorithms," 2020.

[41] K.-S. Na, S.-E. Cho, J. P. Hong, *et al.*, "Association between personality traits and suicidality by age groups in a nationally representative korean sample," *Medicine*, vol. 99, no. 16, 2020.

[42] D. Ramırez-Cifuentes, A. Freire, R. Baeza-Yates, *et al.*, "Detection of suicidal ideation on social media: Multimodal, relational, and behavioral analysis," *Journal of medical internet research*, vol. 22, no. 7, e17758, 2020.

[43] A. O. R. Rodrıguez, M. A. Riaño, P. A. G. Garcıa, C. E. M. Marın, R. G. Crespo, and X. Wu, "Emotional characterization of children through a learning environment using learning analytics and ar-sandbox," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, pp. 5353–5367, 2020.

[44] M. Z. Asghar, F. Subhan, H. Ahmad, *et al.*, "Senti-esystem: A sentiment-based esystem-using hybridized fuzzy and deep neural network for measuring customer satisfaction," *Software: Practice and Experience*, vol. 51, no. 3, pp. 571–594, 2021.

[45] A. K. Kushwaha, A. K. Kar, and Y. K. Dwivedi, "Applications of big data in emerging management disciplines: A literature review using text mining," *International Journal of Information Management Data Insights*, vol. 1, no. 2, p. 100 017, 2021.

[46] S. Datta, P. Mitra, and P. Kundu, "Fervency: A squashy intrigue to ascertain emotions using textual categorization," in *2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA)*, IEEE, 2022, pp. 1348–1355.

[47] S. Gangbo and G. Shidaganti, "Classification of student mental health prediction using lstm," in *2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT)*, IEEE, 2022, pp. 1–6.

[48] S. Ji, X. Li, Z. Huang, and E. Cambria, "Suicidal ideation and mental disorder detection with attentive relation networks," *Neural Computing and Applications*, vol. 34, no. 13, pp. 10 309–10 319, 2022.

[49] S. Dobilas, *Long short-term memory networks (lstm)- simply explained!* May 2023. [Online]. Available: https://databasecamp.de/en/ml/lstms.