Finding the Presence of Outlier Personalities within the Student Community by Applying BERT on Text Inputs

Thesis Submitted in Partial Fulfilment of the
Requirements for the degree of
Master of Computer Science & Engineering

By
Aritra Podder

Examination Roll Number: M4CSE22022
Class Roll Number: 002010502022
Registration Number: 154146 of 2020-2021

Under the guidance of
Dr. CHITRITA CHAUDHURI
Associate Professor

Department of Computer Science and Engineering
Faculty of Engineering and Technology
Jadavpur University
Kolkata – 700032, India
August 2022

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

To Whom It May Concern

I hereby forward the thesis entitled "**Finding the Presence of Outlier Personalities within the Student Community by Applying BERT on Text Inputs**" prepared by **Aritra Podder** (University Registration No: **154146** of **2020-2021**, Examination Roll No: **M4CSE22022**) under my guidance and supervision. It is a bona-fide piece of work that may be accepted in partial fulfilment of the requirement for awarding the degree of **Master of Computer Science and Engineering** in the Faculty of Engineering and Technology, Jadavpur University, Kolkata.

.....

Dr. Chitrita Chaudhuri (Thesis Supervisor)
Associate Professor
Department of Computer Science and Engineering
Jadavpur University, Kolkata-32

Countersigned

.....

Prof. Nandini Mukhopadhyay
Head, Department of Computer Science and Engineering,
Jadavpur University, Kolkata-32.

.....

Prof. Chandan Mazumdar
Dean, Faculty of Engineering and Technology,
Jadavpur University, Kolkata - 32

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

Certificate of Approval*

This is to certify that the thesis entitled "**Finding the Presence of Outlier Personalities within the Student Community by Applying BERT on Text Inputs**" is hereby approved as a creditable study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that, by this approval, the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve the thesis only for the purpose for which it has been submitted.

.....

Signature of Examiner 1:

Date:

Final Examination for evaluation of the thesis.

.....

Signature of Examiner 2:

Date:

*Only in case the thesis is approved

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

Declaration of Originality and Compliance of Academic Ethics

I hereby declare that this thesis entitled "**Finding the Presence of Outlier Personalities within the Student Community by Applying BERT on Text Inputs**" contains literature survey and original research work by the undersigned candidate, as part of his Degree in Master of Computer Science and Engineering.

All information has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name: Aritra Podder

Registration No: 154146 of 2020-2021

Exam Roll No: M4CSE22022

Thesis Title: **Finding the Presence of Outlier Personalities within the Student Community by Applying BERT on Text Inputs**

.....

Signature with Date

Acknowledgement

The satisfaction and euphoria that accompanies the successful completion of this task would be incomplete without the mention of the people who made it possible. Their constant guidance and encouragement crowned my effort with success.

It is a great pleasure to express my sincerest thanks to my project supervisor Dr.Chitrita Chaudhuri, Associate Professor, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jadavpur University, for her encouragement, valuable suggestion, and constant support during the course of this work.

I would like to thank all the professors of the Department of Computer Science and Engineering, Jadavpur University, Kolkata for the guidance they provided me throughout the duration of the Master of Computer Science and Engineering course.

A special note of thanks goes to Prof. Nandini Mukhopadhyay, Head, Department of Computer Science and Engineering, Jadavpur University.

I would like express my gratitude to Prof. Chandan Mazumdar, Dean, Faculty of Engineering and Technology, for providing an excellent environment for the completion of this work.

I am also indebted to my co-researcher Mr. Anupam Baidya for his seamless co-operation, help and inspiration during the period of research. I am thankful to my fellow classmates and my family too for their constant help and support.

.....

Signature and Date

Name: Aritra Podder

Registration No: 154146 of 2020-2021

Exam Roll No: M4CSE22022

Index

Chapters **Pages**

Index (contd.)

List of Tables

List of Figures

Chapter 1: Introduction

COVID-19 pandemic is the current source of worry all across the world. In early December 2019, an outbreak of coronavirus disease 2019 (COVID-19), caused by a novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), occurred in Wuhan City, China. On January 30, 2020 the World Health Organization (WHO) declared the outbreak as a Public Health Emergency of International Concern. Perceived risk of acquiring disease has led many governments to institute a variety of control measures. For this a sudden Lockdown had been implemented in India which led to an overall unstable situation. All aspects of public life including education and economic factors got affected in the process.

It has been felt that the students suffered the more because they are at an impressionable age. Many of them had to vacate their hostels on an emergency basis, and were forced to return to their homes due to a sudden closure of their Institutes. Disruption of all academic activities for a prolonged period left a great impact on the mental health of such students. There followed a time during which most schools and colleges reverted to online mode classes. The burden of ill-organized online courses, started haphazardly soon afterwards, added to the chaotic situation. This also resulted in a sharp digital divide amongst the community, as not all of them could afford the luxury of undisturbed internet connections required for such online mode of education. Final year students lost job opportunities for not being able to avail the online recruitment process. Moreover, some even faced financial crisis due to loss of job or business of their earning family members during the lockdown. Other collateral damages included COVID infections and emergency hospitalizations, culminating to death in many cases.

After a huge protest educational institutes have been reopened, but the ill-effects of the dark times may have resulted in an immensely handicapped society. The present work attempts to capture the resulting aberrations within a small community of university students. A survey was conducted with questions regarding general issues related to the pandemic, as well as those involving government policies on education and economy. The responses received served as textual inputs to an integrated automated system which helps to decipher outlier mentalities prevailing amongst the students in these disturbed times.

Many works have been done on sentiment analysis where the sentiment of an author is designated as either positive or negative [1]. Mostly the social media data is used in case of sentiment analysis as it is open and vast, and people are free to share their opinion in such media [2]. Some of these use twitter data by finding textual information and diffusion patterns from tweets and performing sentiment analysis based on interactions [3]. Combination of such different features have been used in many works. The negative aspect of the data utilized is that one cannot trace the persons behind the culprit tweets.

Thus, in this research, the work is done on a closed group of students within the university. The data consists of textual responses to twelve online survey-based questions related to the pandemic times – six involving education, one involving economy and five involving general topics. Moreover, instead of relying on bare positive and negative sentiment values, the work is based on extracting emotion-based attribute trio Valence-Arousal-Dominance which is felt to measure the trend of personality bias more effectively. The strength of Deep Learning Model BERT [4] in deciphering the contextual semantics of expressed words is also explored in detail.

1.1 Scope of the Work

This work has been done as a part of a year-long Master's degree dissertation in the domain of Computer Science and Engineering. The research involves several areas in the domain – e.g., Natural Language Processing, Machine Learning and Deep Learning techniques.

The objective of the work is to find out students who are mentally alienated from the rest of their batchmates, as detected from their involuntary usage of words and expressions.

Many state-of-the-art tools and concepts have been utilised to fulfil the tasks satisfactorily. The researcher feels that the results obtained using the proposed machinery are promising enough to justify the efforts spent therein.

1.2 Motivation

The motivation behind the work is the growing need to detect persons who are becoming detached from the mainstream society. The current pandemic-induced scenario provides a fostering ground for the growth of such negative mentality. Unless some prompt action is taken to counsel the wayward personalities back to normal spheres, heavy losses may occur all around. Already the footsteps of such disasters are heralded by the recent growth of tragic affairs being reported daily in the news media.

The proposed system can help in delivering the alerts in advance and in a totally non-invasive manner, as the topics asking for opinion within the involved group are in themselves harmless and justified under the circumstance. It is the thought-provoking nature of the subjects that help to ease out the pent-up emotions of the participants, but that too in an entirely non-judgemental manner. With the help of such work, one can hope to track and alert the machineries beforehand and thus pave the way towards a better tomorrow for the affected personalities.

1.3 Thesis Layout

The present Chapter 1 provides a brief introduction to the thesis and shortly discusses its utility in recent times.

In Chapter 2 are introduced some past works related to the present one. Some of these directly or indirectly influence the current study.

Chapter 3 describes the background details and theories involved in the proposed work.

The actual architectural design of the system, along with the algorithms required to implement the task has been detailed in Chapter 4 of this thesis.

Chapter 5 first discusses the data and major tools involved in the process. Then it goes on to incorporate the results and inference obtained from the system in a comprehensive manner.

In Chapter 6 are inscribed the overall conclusions drawn. Some promising future scopes are also hinted at here.

The citations of previous works are supported by a Reference section at the end.

Chapter 2: Previous Research Work

This chapter surveys previous research works on topics related to emotion detection and natural language processing which have got direct connection with the current work.

Applying the machine in tasks which humans performed so far is an age-old practice. NLP have been a lucrative choice for many of the early researchers in this context. This is amply instanced by Osgood et al. (1957) [5], who had delved to find meanings from words using semantic differential in a general way. Attitudes, effects of psychotherapy, cultural and language meaning differences, and assessment of personalities were used in this research.

Soon, Computer Scientists started looking for new attributes which can extract the essential sentiment value of words in a more comprehensive way. To this end, Russell (1980, 2003) [6, 7] performed analysis of emotion words in two separate studies. In the first, he introduced the idea of three primary independent dimensions of emotions - Valence or pleasure, Arousal and Dominance. In the next paper, he demonstrated how fine-tuned emotions such as joy, anger and fear can be plotted within a three-dimensional space of valence, arousal, and dominance. These pioneering works on the said emotion norms were soon to be followed up by many more enthusiasts, as is evident from the researches detailed next.

Bardley and Lang, 1999 [8] in their experiments, asked participants to rate more than thousand English words using affective dimensions of Valence, Arousal and Dominance. This heavily contributed towards the emotional ratings offered in their Affective Norms of English Words (ANEW). Warriner et al., 2013 [9] created a lexicon similar to ANEW but with 14000 words in all that provided a richer source of emotional information. Mohammad, 2018 [10] created another database of 20000 English words expressed in terms of their valence, arousal and dominance scores, and named it as NRC-VAD Lexicon. It is one of the largest manually created VAD lexicon with the help of best-worst scaling.

These resources are used by other recent works. For instance, Wu et al., 2019 [11] attempt to predict VAD scores from sentences based on variational autoencoders. On the other hand, Zhu et al., 2019 [12] use adversarial learning, and Akhtar et al., 2019 [13] utilise ensemble framework for the same purpose. There exist VAD lexicons in languages other than English as well, such as the ones created by Moors et al., 2013 [14] for Dutch, by Vo et al., 2009 [15] for German, and by Redondo et al., 2007 [16] for Spanish.

Meanwhile, in a separate vein, the search for the simplest sentiment expressions in the form of positive and negative aspects of a word, have been on-going for a long time. Andrea et al, 2006 [17] created SentiWordNet which is an opinion lexicon adapted from WordNet database. The SentiWordNet finds positive, negative and object score of words in English. Baccianella et al [18] discusses usage of SentiWordNet for classification of sentiment and opinion mining, by assigning a positivity, negativity, and a neutrality score for text.

Going one step further, Shailendra Kumar Singh and Sanchita Paul, 2015 [19] stated that in the sentiment analysis process, negation words and negative prefixes play an important role in sentiment score calculation, as they can reverse the sentiment of sentences. The opinion words and opinion phrases are used to extract positive or negative sentiments. So, new horizons continued to be opened by erstwhile researchers all the time.

Apart from emotions, another interesting task associated with human text is the process of summarization that needs to be necessarily applied to large corpuses. Transferring the brunt of the job to the machine was just one step forward. H.P. Luan et al, 1958 [20] proposed a model that could find the abstract of a technical literature conveniently. The relative significance of a word or a sentence was derived by using statistical information such as word frequency and the distribution of each word in a particular sentence. Sentences with the highest scores were to be included in the abstract.

Over the years several more researches were attempted in this domain. J.L. Neto et al, 2002 [21] built a model of automatic text summarization that used Naïve Bayes and C4.5. C4.5 is a machine learning algorithm utilised in data mining which generates *Decision Tree Classifier* developed by J. Ross Quinlan, 1993 [22]. R. Barzilay et al, 2005 [23] used a Sentence Fusion technique, which involved finding of similar information across documents to summarize new articles in the abstractive mode. R. Khan et al, 2019 [24] have described a system of Extractive summarization using K-Means clustering and TF-IDF (Term Frequency and Inverse Document Frequency) to calculate the overall weight of each sentence and generate the summary.

Research progressed in search of better avenues to attain the goal in Natural Language Processing. Jacob Devlin et al [4] in 2018 introduced a new language representation model called BERT, designed to pretrain deep bidirectional sequences from unlabelled text by jointly conditioning on both left and right context in all layers. It has been found that this model can be fine-tuned to create state-of-the-art systems for a wide range of tasks. In 2019 Yinhan et al [25] added some enhancements on BERT. These enhancements include longer training sessions, larger batches with more data, and lengthier sequences. It incorporates dynamic mask patterns while training and also omits next-sentence-prediction. A new dataset, CC-NEWS is being used on the model.

In the present work, certain other areas of machine learning needed to be touched upon. One of these involved comparing similarity between entities based on text content. The following research endeavours contributed towards establishing the state-of-the-art on such aspects.

Alfirna Rizqi Lahitani et al, 2016 [26] studied and computed similarity measurement of online documents by considering their tf-idf scores. Vectors formed from these scores represented the documents and they were compared with the help of cosine similarity measurements. Similarity can also be found by cluster analysis techniques. In 1996 Martin et al [27] presented a new clustering technique DBSCAN, which relies on a density-based notion. This technique is able to produce arbitrarily shaped clusters containing similar items, and singleton elements marked as outliers. In 2020, Dingsheng Deng [28] explored the DBSCAN clustering technique and tried to improve the technique by using big data processing.

Armed with all the experiences gathered so far, it is time to touch upon the exact contexts required for the present work. These are presented in the following chapter.

Chapter 3: Background Studies

This chapter addresses the details of some basic topics utilized in this research work. The first foundational topic to be discussed is Machine Learning.

3.1 Machine Learning

Chief amongst the core subjects explored in this thesis work comes Machine Learning [29]. As the name suggests, machine learning is nothing but the process to train the computer, so that it helps to learn and improve a program for better results. It focuses to develop computer programs that can access data as required and learn from them so that the machine can provide results on some unseen data. Similar to human brain, machine learning depends on inputs for learning and gains knowledge and understanding. These inputs are known as training data. These data help to detect patterns and similarities based on which inferences are drawn that helps to process unseen data. The primary aim is to allow computers to learn autonomously without any human assistance and perform actions accordingly.

The concept of machine learning was coined by Arthur Samuel [30], a computer scientist at IBM. He designed a computer program for playing checkers. The objective of the program was: the more the game would be played, the more it would learn from experience. Armed with this gained experience, the algorithm would make predictions for the next move. Thus, with the help of learning, the program would perform better progressively. Two other obvious advantages of machine learning, are that – firstly, it completes tasks in a much shorter time; and secondly, it removes chances of human errors and bias to a great degree. But again, to achieve these goals, there is need for air-tight and fool-proof algorithms that drive machine learning tools critically towards success. A mathematical model is built based on the sample data or training data that helps to make predictions or decisions without being explicitly programmed to do so. A reliable tool utilised for this purpose is the neural networks, which aims to simulate the biological brain [31].

Machine learning can be utilised for a variety of tasks such as sensing model security vulnerabilities, or high-risk factors in a system, or providing online shopping recommendation, or detecting either health issues, or fraud in multiple domains, and many more such interesting issues.There are mainly four types of learning involved in these tasks:

- Supervised Learning
- Unsupervised Learning
- Semi-supervised Learning
- Reinforcement Learning

1. Supervised Learning [32]: In this type of machine learning labelled data is fed to the machine. By analysing the training dataset with known labels, the learning algorithm produces a derived function to predict output values. The resulting function-based model can then be used to predict the output of some unseen data. This type of learning also helps to compare known labels with predicted results. If accuracy is not found to be satisfying then the model can be modified accordingly. Thus, the main criteria for supervised learning are annotated data labelled with class values, and hence the process is known as Classification. But in this thesis work, the obtained data is not priorly labelled, so supervised learning is not used here explicitly.

2. Unsupervised Learning: This type of learning is in direct contrast to supervised learning. In this technique, unlabelled or unclassified data is fed to the machine. The algorithm tries to infer a similarity measure which establishes meaningful connections within unlabelled data, often forming well defined clusters in the process. Consequently, the technique is also known as Clustering or Cluster Analysis.

3. Semi-Supervised Learning: As the name suggests, this type of learning is a mixture of both supervised and unsupervised learning. Mostly the model is fed with labelled data, which is usually small in number, but may be used to train the model to label unlabelled data. Thereafter iterative techniques of increasing accuracy and efficiency of the model are explored. In the process the knowledge-bank of the system also gets enhanced with a better understanding of the dataset.

4. Reinforcement Learning: This type of learning interacts with the environment by producing actions and discovering errors or rewards. With the help of trial-and-error search and feedback, both the reward as well as the learning is maximized. This helps to automatically determine the ideal behaviour within a specific context, and thus achieves the best performance under the circumstance.

A subset of machine learning is Deep Learning, which is more advanced in nature. It consists of three or more neural networks. A neural network sets up a series of algorithms to recognize the underlying relationship in a set of data through a process that mimics the operation of a human brain. In deep learning, the neural networks simulate the human learning pattern with intrinsic training acquired through large amounts of data. The concept of using multiple neural networks help to optimize and refine the model for better accuracy.

The basic difference between Deep Learning and Machine Learning can be understood with the help of a simple example. Assuming there is a dataset with photos of different pets like cats, dogs, rabbits etc. -with the help of deep learning algorithms, important features to distinguish each animal can be determined (like ears, legs). But basic machine learning usually deals with specific features that needs to be priorly connected to each distinct class manually by a human expert. Further, the deep learning algorithm adjusts and fits itself better to accurately predict the class of an unseen entity.

Deep learning attempts to map to the correct result through a combination of data inputs, weights and bias. It consists of multiple layers of interconnected nodes, where the output of each node is fed as input to the next node in order to refine and optimize the output. This progression of computation is called forward propagation. Except the input and output layer, the middle layers are called hidden layers. There is another process involved, known as backpropagation, which helps to minimize the calculated error after each forward propagation by adjusting the weights and biases of the layers. Multiple forward and backward propagation helps to improve the performance accuracy to a great extent. Deep learning can be used in multiple domains such as speech recognition, computer vision, text summarization, image colouring, fraud detection, language translation, pixel restoration etc., amongst which only a few are utilised here. The following sections describe these select few in more detail.

3.2 Natural Language Processing (NLP)

Every human language is filled with ambiguities which make it difficult for the software to accurately identify the meaning of a text or voice data. Homonyms, homophones, idioms, sarcasm, metaphors, grammars make a sentence structure vary a lot. The irregularities are understandable by humans easily, but the task becomes immensely difficult for the machine. Proper training on huge pre-labelled datasets along with the right choice of machine learning techniques could lead to the correct output for any tasks involving text inputs.

NLP [33] is a specialized field within Machine Learning which helps the machine to understand, analyse, manipulate and generate human language. NLP generally utilises several statistical, machine learning and deep learning techniques. NLP is prolifically used in tasks such as translating text from one language to another, summarizing texts etc., and comparing the semantic content of two text inputs, to mention just a few. One of the most common works that comes under NLP is Sentiment Analysis, which is explained in the following section.

3.3 Sentiment Analysis

Sentiment analysis is the process of identifying sentiment or emotion lying in a text. It is also referred as opinion mining. The measure provides helpful insights into how audiences or spectators sense and understand a topic, or how an event or text impacts the thinking of a listener or viewer. This helps to get an idea about the quality or success of a newly launched product, scheme or event, by assessing its impact on people through the words of texts expressed in favour or against. Generally social media platforms are utilized for gathering the sentiment of the populace.

The most general type of sentiment categorisation is expressed in the form of positive and negative sentiments corresponding to expressions for or against the topic or item. A third emotion may also be envisaged to indicate non-committal or neutral sentiment. In NLP schemes words involving such emotions are usually pre-labelled with numeric values corresponding to both positive and negative sentiments. The range of values may vary, typical limits being expressed through real numbers between -1 and +1, with -1 representing the most negative sentiment and +1 the most positive sentiment. In this scheme 0 represents neutral sentiment. However, in the scheme followed in this research, the prevailing values used for most negative and most positive sentiments lie between 0 and 1 (with 0.5 indicating neutral value) according to the dictums of SentWordNet Lexicon [18].

Although this type of sentiment analysis, more generally referred to as Opinion Mining, suffices in most cases, there may be occasions where a more detailed analysis of the state of the human mind regarding some incident or environment is needed. Such a probe demands other measures of assessments. One such involves three critical metrices – the Valence-Arousal-Dominance scheme – details of which is discussed next.

3.4 Valence-Arousal-Dominance (VAD)

Emotions are part and parcel of a human's life. They express one's mental reaction towards a stimulating event with physiological, behavioural indications – one such being words used in the context. These can help to detect mental illness or mental distortion or depression by combining sets of features extracted from those words. If such detection can be done at an early stage, people can undergo proper preventive measures to avoid future catastrophic outcomes later in life.

Valence-Arousal-Dominance, a three-dimensional model, is a metric by which the emotional involvement of a person can be evaluated from the text words used by the person. This technique was introduced by Mehrabian and Russell in 1977 [34]. VAD is emotion-based sentiment analysis. Information about affective meaning of words is used to find emotions and moods. The VAD model consists of three components, valence, arousal, dominance. *Valence* is defined by the pleasantness of a stimulus; *arousal* is the intensity of emotion created by a stimulus and *dominance* is the degree of control by a stimulus. *Valence* basically indicates the positivity or negativity, in other words the measure of happiness or unhappiness imposed by a stimulus; *arousal* is how much the stimulus excites or arouse a person; and *dominance* indicates how much the stimulus dominates the person or controls the person. All three are measured in a scale between 0 and 1 in the present research. For valence 0 indicates the extreme limit of unhappiness, while 1 corresponds to the happiest condition. For arousal 0 indicates the state of being least aroused and 1 indicates the state of being most aroused. For dominance, on the other hand, 0 refers to the most controlled or least dominating personality, whereas 1 indicates the most dominating personality.

The visual effect of the VAD attributes is presented next for a sample text input in figure 3.1 below. Here, the presence of outlier data elements with respect to the three attributes individually and collectively is clearly evident from the 3D view.

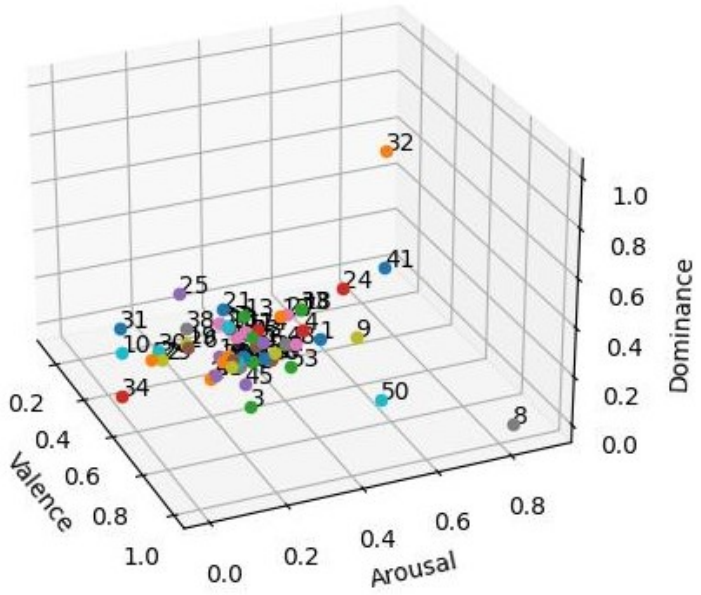

Fig 3.1: 3D View of a sample VAD model

In the present research work, the first two features, namely valence and arousal are calculated from sentiment scores indicated by SentiWordNet [18]. However, for dominance values a deep learning approach involving BERT (Bidirectional Encoder Representations from Transformers) has been utilised. The next section discusses the topic in more detail.

3.5 BERT

Advanced NLP combines both machine learning and deep learning to extract, label and classify text. An example of such framework is BERT. BERT is an open-source machine learning framework for NLP. It attempts to extract the meaning of a text with the help of surrounding texts.

BERT [4] is a transformer used to overcome the limitations of neural networks such as in the case of long-term dependencies. It is a pre-trained model which is bidirectional and can be tuned as per requirement to perform several NLP tasks. BERT models are pre-trained on huge datasets - thus no further training is required during the application. It uses a powerful architecture with inter sentence transform layers so as to get the best results in tasks such as summarization. The working principle of BERT is elaborated in figure 3.2 below.

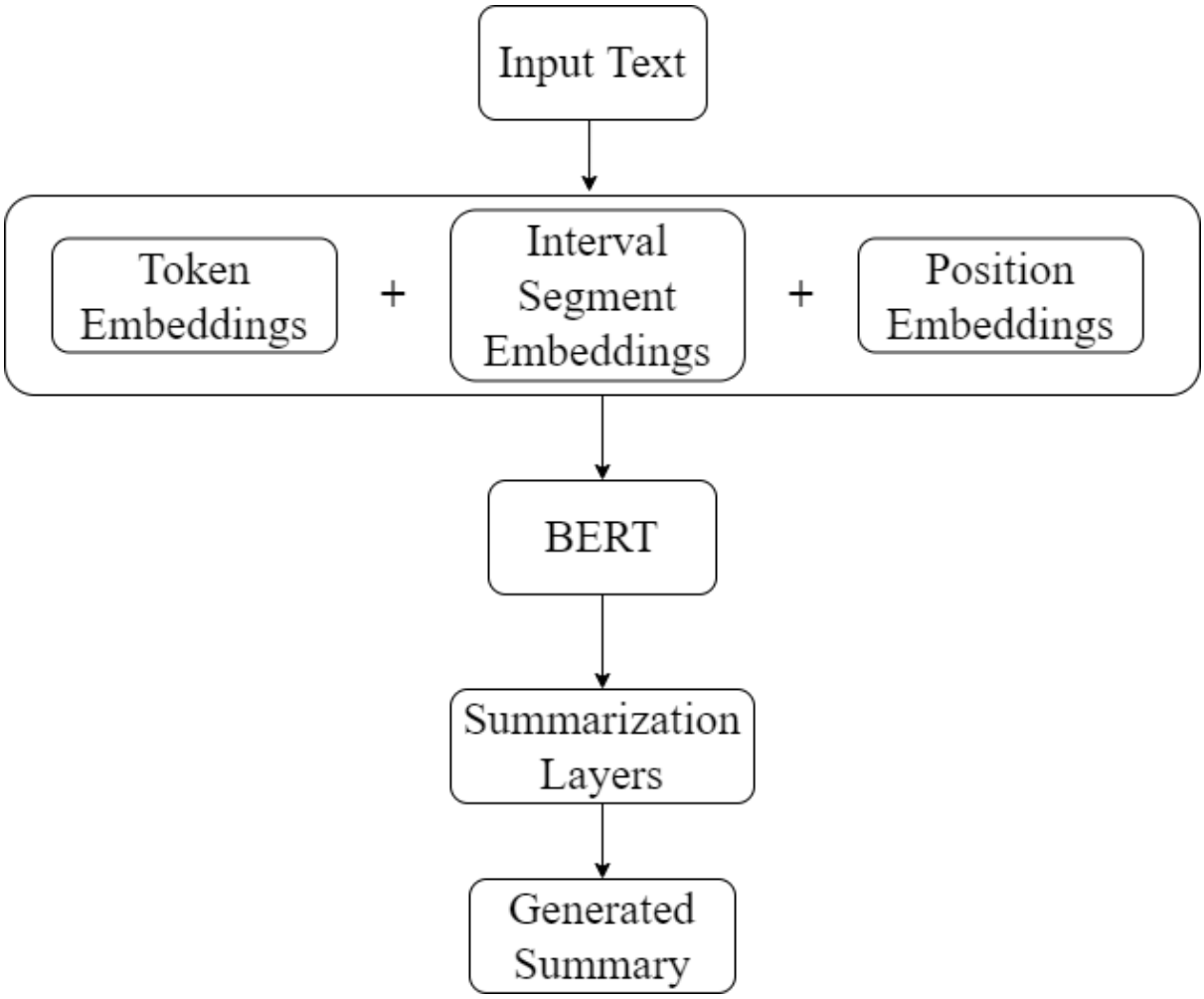

Fig 3.2: BERT Methodology flowchart [35]

BERT is trained as a masked model where the output vectors are tokened. It makes use of embeddings for indicating different sentences and it has only two labels, namely sentence A and sentence B rather than multiple sentences. According to required summary, these embeddings are modified. The technique is illustrated in the following figure 3.3.

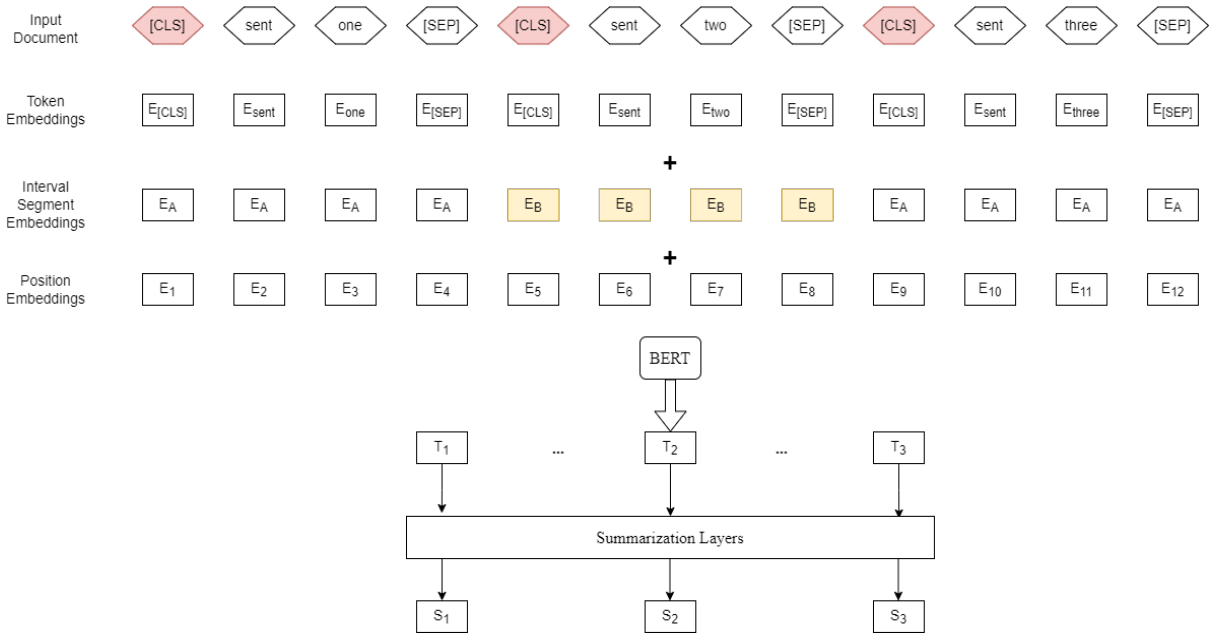

Fig 3.3: Overview of BERT [35]

Initially each sentence is pre-processed with prefix CLS tag and suffix SEP tag. The CLS tag is used to aggregate the features of one or more sentences. In interval segment embeddings, sentences are assigned label as E_A or E_B alternatively. This is done to distinguish sentences in a document. Embeddings basically refers to the representation of words in their vector forms, which helps to make their usage flexible and compatible for mathematical calculations. Even the Google utilizes this feature of BERT for better understanding of queries. Embeddings helps to develop a similarity model between the words. The vectorization of the words is represented in the following figure 3.4.

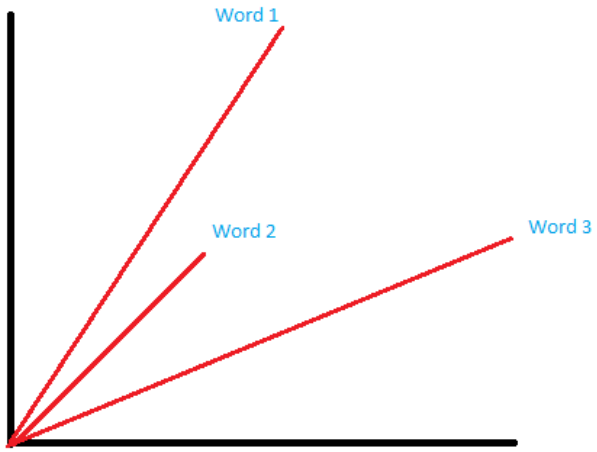

Fig 3.4: Vector representation of different words

Three types of embeddings are applied to the text before feeding it to the BERT layer:-

1. Token embeddings, where words are converted into a fixed dimension vector.
2. Segment embeddings, which is used to distinguish different inputs.
3. Position embeddings where BERT can support input sequences of 512. For input sequence of 512, the resulting vector dimensions will be (512,768). Every word is stored as a 768-dimensional representation. Positional embedding is used because alteration of the position of a word in a sentence may alter the contextual meaning of the sentence and thus should not have same representation as vectors.

There are two types of BERT models available. One is BERT Base which has 12 transformer layers along with 12 attention layers and 110 million parameters. Another one is BERT Large which consists of 24 transformer layers along with 16 attention layers and 340 million parameters. In this thesis work, BERT Base model is being used. Transformer layer is actually a combination of encoder and decoder layers along with intermediate connections. Each encoder includes Attention Layers along with an RNN. Decoder also has the same architecture but it includes another attention layer in between them which helps to concentrate on important words. Following figure 3.5 presents the scenario concisely.

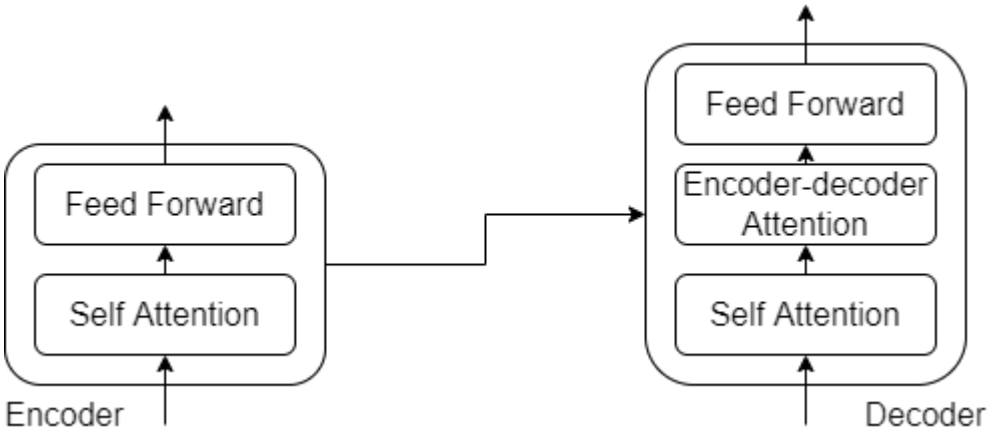

Fig 3.5: BERT Architecture [35]

BERT model tries to identify the strongest links between the words with the help of self-attention layer. For summarization, different types of layers can exist. A simple classifier or a linear layer is added to the BERT along with a sigmoid function to predict the score. In case of inter sentence transformer, various transformer layers are added to the model which makes the transformation more efficient. In recurrent neural network, an LSTM (Long Short Term Memory) layer is added with the BERT output in order to learn summarization features. BERT is already present in the python library. One just needs to install it and import the required packages and model. BERT extractive summarizer is used here to find gist of text responses from individuals.

3.5.1 Summarizer

Summarizer is a predefined model in BERT. It is used to summarize a text. Summarizer helps to shorten a text while keeping the original meaning intact. Two types of summarization processes are there as represented in figure 3.6 next.

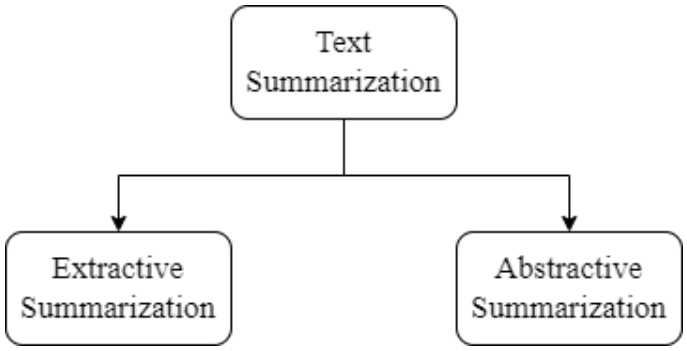

Fig 3.6: Types of Summarizers

In Extractive summarization the important phrases or sentences (top n) are identified from the original text and extracted. These extracted sentences together form the summary for the original text as illustrated in figure 3.7.

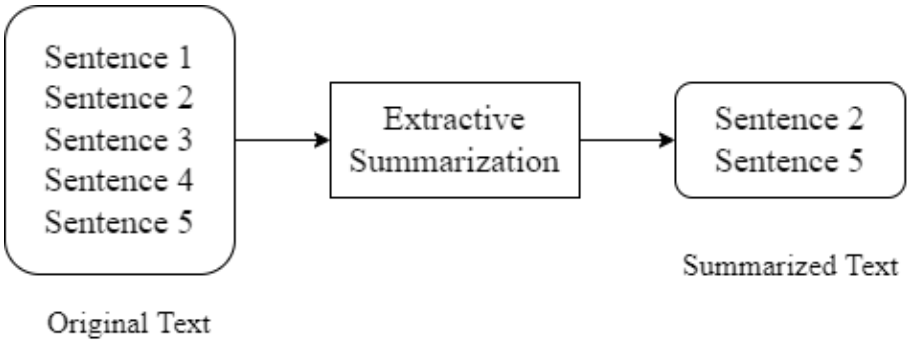

Fig 3.7: Extractive Summarization [36]

On the other hand, in abstractive summarization, new sentences are generated from the original text. The newly generated sentence may not be present in the original text. The process is depicted in figure 3.8 below.

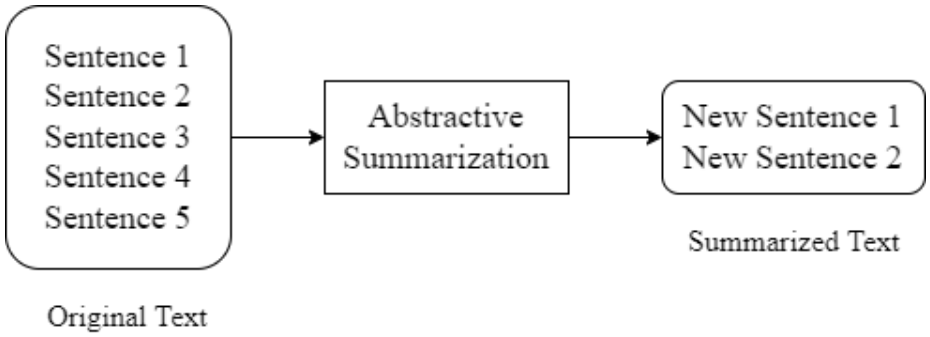

Fig 3.8: Abstractive Summarization [36]

The summarizer takes the text as a mandatory parameter. Optional parameters like *number of sentences* can be supplied as a ratio or an integer.

3.5.2 Sentence Transformer

It is difficult for the machine learning algorithms to analyse a raw corpus. One needs to convert such data into a numerical representation best expressed in the form of a vector. *Sentence Transformer* is the framework which encode the individual sentences into vectors, making it easier to analyse or consume that text by a machine learning algorithm. Then the model can apply mathematical operations and get insights from the data.

Sentence Transformer is a python framework used for sentence, text, image embeddings. These embeddings can thereafter be compared with some similarity function such as the *cosine-similarity*, to find sentences having similar meaning. More than 100 languages are supported by the sentence transformer. It maps sentences and paragraphs to a 768-dimensional dense vector space.

The sentence transformer in python can be loaded by importing pre-trained models, and passing the model's name as follows: SentenceTransformer('model-name'). In this thesis work, '*bert-base-nli-mean-tokens*' is used as the model for sentence transformer.

3.6 Cosine Similarity

The sentence transformer helps to convert the sentences into vector form. These vectors can thereafter be compared to find similarity between the sentences. One of the most popular similarity finding technique is *Cosine Similarity* [26]. It measures the similarity between two non-zero vectors belonging to an inner product space, by expressing the same as the inner product operation on them. It measures the cosine of the angle between them. The following formula calculates the cosine similarity between two vectors A and B:

$$\text{Similarity} = (A.B) / (||A||.||B||)$$

In sklearn module of python, an in-built function is available to calculate the cosine similarity. The function accepts two vectors of similar dimension as parameters and returns a similarity value between them in the range 0 to 1. Cosine similarity is preferred over Euclidean distance in this instance as two documents represented by the Euclidean technique can be far apart due to their textual size difference. But the cosine similarity will try to find the angular distance between them, as depicted in the following figure 3.9.

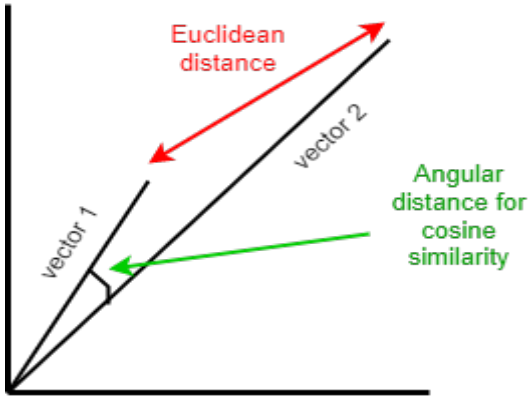

Fig 3.9: Exploring cosine similarity

From the foregoing illustration, it is clear that vectors 1 and 2 are closely placed i.e., similar. But due to their textual size difference, the Euclidean distance is high, giving rise to a contradictory conclusion of being dissimilar. The closer the two vectors, the lesser is the angular distance. Thus, more the similarity between two vectors, the cosine similarity value will be closer to 1.

3.7 Clustering

Clustering basically refers to grouping of data. Depending on feature(s), a set of data can be grouped so that similar type of data will be in the same group. All data mapped to a single cluster will usually be closely knit within the feature space. On the other hand, features of data from different clusters will differ a lot or be far apart.

Clustering is used to group unlabelled data. It is used as a process to find new aspects such as meaningful structure within the data, explanatory underlying processes, or generative features. With the help of clustering, one can discover patterns, outliers, and many more hitherto hidden factors.

For example, consider a set of data points plotted as in Fig. 3.10 below:

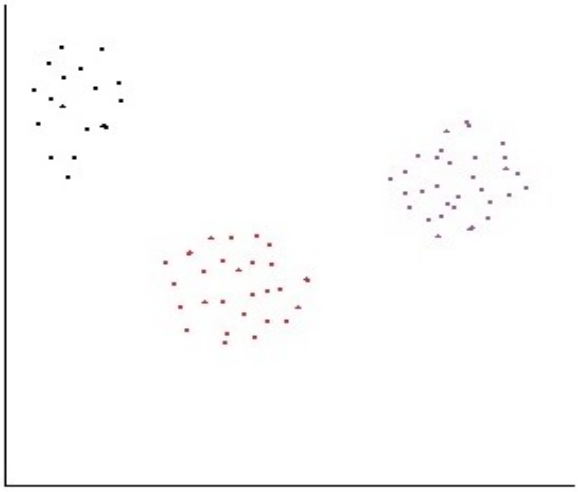

Fig. 3.10: Unlabelled plotted data points

Clustering of the above points can be done as follows in Fig. 3.11 below. Proximus points are taken under the same group/cluster. First the similarities are calculated and then the clustering is performed.

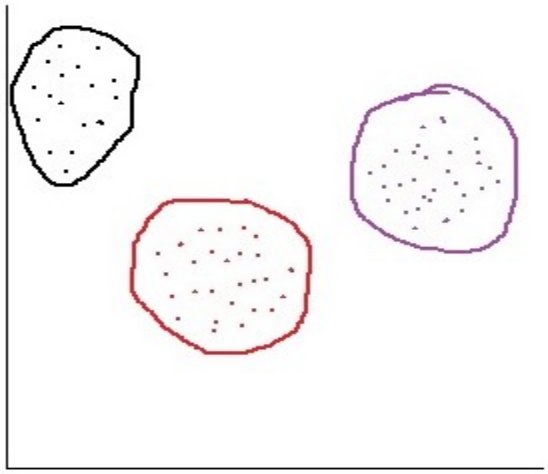

Fig 3.11: Clustering of unlabelled data points

One of the most popular Clustering algorithm is DBSCAN (Density-Based Spatial Clustering of Application with Noise) algorithm [27]. This algorithm clusters the points by measuring the distance between them. The key idea for DBSCAN is to maintain the following rules - each point of a cluster should lie within an *epsilon* distance from its neighbours within the same cluster, and there should be at least a minimum number of such neighbours for the internal points or *core* points in the cluster.

Thus, two parameters are required for clusters to form with this algorithm - the epsilon distance, and the minimum number of points within that distance. If epsilon distance is too small then a large part of data will not be clustered and then the clustering algorithm will result in too many outliers. The epsilon distance varies from application to application. The minimum number of points often depends on the size of the dataset. If dataset is large then minimum number is also chosen to be higher.

This algorithm involves three types of data points. Core points are the points which have more than minimum number of neighbours within epsilon distance. Border points are the points which has less than minimum number of points as its neighbour within epsilon distance but a core is its neighbour. All other points which are not core or border are known to be noise or outliers.

3.7.1 DBSCAN Algorithm:

1. Arbitrarily pick a point from the dataset and mark it as visited.
2. If the point has minimum number of points as its neighbour within epsilon distance, then add all these points to a cluster, after first marking them as visited too.
3. Recursively iterate through the newly added neighbours in turn, and find new points within their epsilon distance. Add these new points to the same cluster as well, after first marking them as visited too.
4. Iterate through steps 1 to 3 for all the remaining unvisited points.
5. When all points are marked as visited, if any are found to be not added to any cluster, then those are marked as outliers or noise.

The next chapter elaborates the design and algorithmic base of the proposed system.

Chapter 4: Methodology

This chapter describes in detail how the objective of the research work is being achieved.

As the goal is to categorise students on the basis of emotion displayed through usage of words, textual responses of university-level students, to 12 questions on three different topics, are collected via an online survey. All questions are related to the impact of COVID on society during the past 2 years. Their words indicate how well or poorly the students are coping with the associated problems. As already pointed out, this is a comparatively easy task for a psychologist, or even a layman – but not at all easy for the machine to assess on its own. For achieving any degree of success, the machine needs to rely on the expertise of NLP tasks and associated NLTK or Natural Language ToolKit [37]. The Python Programming language is especially suited for the purpose, as it provides a wide range of tools and libraries for extracting functional resources to implement NLP tasks on textual inputs - such as stemming, tokenization, parsing etc.

When a user is posed with a question, he/she may respond either elaborately or very shortly. Now, the portion of the answer which actually contains useful data may thus be meagre indeed! However, technological advancements have reached a point where machine algorithms can help to analyse these responses. Again, the responses, when combined on all topics for each student individually, or agglomerated topic-wise for all students, may result in too long a content in some cases. To facilitate the process, the long texts have to be first summarized to meaningful gists. With the help of Machine Learning (or Deep Learning) one can extract the gist of the individual topic-wise or collective responses. Text Summarization can be achieved by using “bert-extractive-summarizer”. It helps to reduce the text bodies while keeping their original meaning intact or by giving insights into their original content.

Next, to extract emotion using VAD values, two structural concepts are in force. The first of these utilise SentiWordNet sentiment values of individual words to evaluate Valence and Arousal from text. The second evaluates Dominance of each candidate using BERT for extracting the semantic essence of the individual responses. For the ultimate evaluation of similar category students, as well as outlier category students, these calculated values have to be converted into numeric vectors using deep learning techniques. So, such learning techniques embedded in BERT are also being used here to measure Cosine Similarity between vectors representing text inputs. These Similarity markers help to derive the Dissimilarity or Distance measures between individuals’ VAD values, which then allows clustering techniques such as DBSCAN to be applied on them to detect related clusters and outliers.

The architectural flow of the current research work is discussed next in the following sections.

4.1 Overall Workflow

There are three separate systems which are explored in the current research:

System 1: This is the Proposed System which calculates VAD values using ML/DL techniques for assessing the mental conditions of each of the participants in the present study, and outputs the same in the form of V, A and D vectors. The first four 4 modules within this system are elaborated in consequent sections 4.2, 4.3, 4.4, and 4.5.

The output VAD vectors are utilised in two separate modules – the first being Module 8 expressed in section 4.9 to find clusters containing similar participants and outlier ones too using DBSCAN algorithm; the other a comparison model for Systems 1 and 2 outputs, as illustrated in Module 9 which is discussed in section 4.10.

System 2: This one deals with the Benchmarking System which is designated as the NRC System of calculating VAD values using the NRC Lexicon [10]. There is a single Module 5 in this system which is described in section 4.6 next. This system's output vectors are utilised in modules elaborated in sections 4.9 and 4.10 as already explained above.

System 3: This system uses the raw text lines and applies DL techniques on these initially to generate summarized gist response vectors for each participant. This action is represented by Module 6 described in section 4.7. The participants' response vectors are then compared using Cosine Similarity to achieve a reversed distance parameter between them as explained in Module 7 discussed in section 4.8. This module further appropriates the clustering scheme of DBSCAN to assess content-based outlier personalities, if any, from raw texts.

Ultimately, the outputs of the three systems are collectively assessed in module 10 as described in section 4.11.

The complete scheme is illustrated in the following figure 4.1.

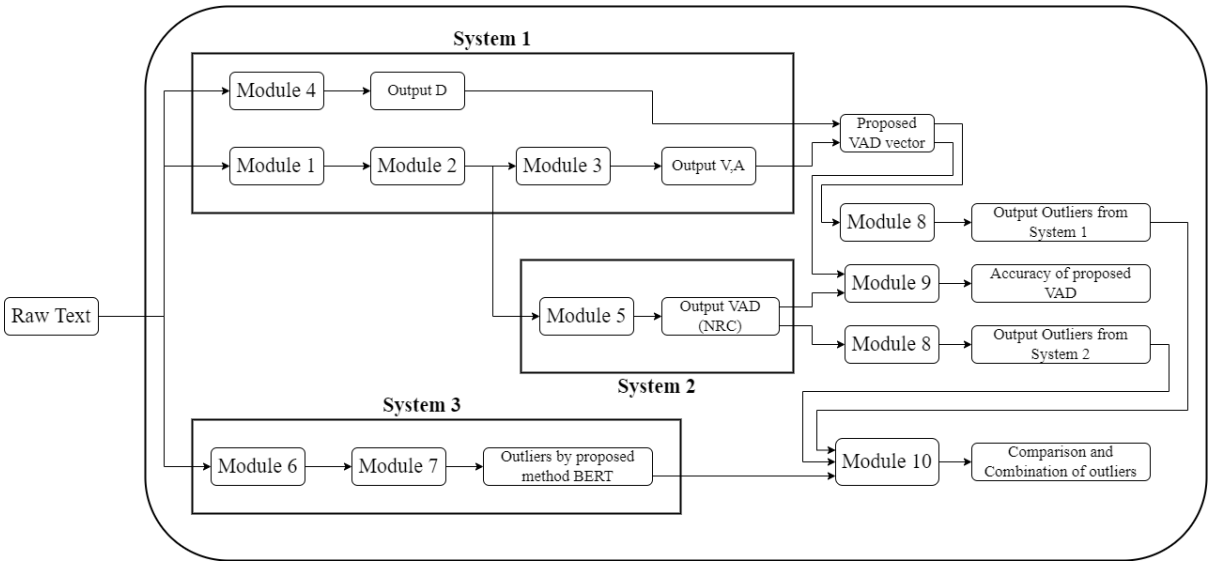

Fig 4.1: Overall Workflow Diagram

4.2 Module 1: Pre-processing

Input: Raw Text

Output: Cleaned Text

Module 1 takes the raw input and perform cleaning on the input by removing redundant data. In other words, data pre-processing is done in Module 1.

Algorithm followed for Data Pre-Processing using NLTK

1. All the required columns are processed to lowercase by using regex.
2. Numbers were removed from the responses by using regex
3. Website URLs (phrases containing 'https') are removed by using regex
4. The cleaning process is performed on the four columns (one containing combined responses, one containing education responses, one containing economic responses and last one containing general responses).

4.3 Module 2: Tokenization

Input: Cleaned Text

Output: Parsed Text Words

In this module, tokenization is performed, which helps to find sentiment value of words from SentiWordNet.

Word Tokenization Scheme using NLTK:
In this phase all words, except the stopwords, are tokenized or marked as adjectives, verbs, nouns and adverbs. This helps at a later stage to get the sentiment score of each word in its correct sense from SentiWordNet. The tokenization is done on all four columns of the dataset - combined responses, educational responses, economic responses and general responses.

After tokenization of the responses, the responses are lemmatized to verify if the tokenization is done correctly or not.

The whole process of parsing is captured in a sample snapshot in Figure 4.2 below.

```
word_df
```

	all_words	Cleaned Survey	Cleaned Survey edu	Cleaned Survey eco	Cleaned Survey gen	POS tagged	POS tagged edu	POS tagged eco	POS tagged gen	Lemma
0	Practical knowledge intake has become less.,On...	Practical knowledge intake has become less Onl...	Practical knowledge intake has become less Onl...	Just the transport expenses are being saved no...	Basics norms set by government ate followed So...	[(Practical, n), (knowledge, n), (intake, n), ...	[(Practical, n), (knowledge, n), (intake, n), ...	[(transport, n), (expenses, n), (saved, v), (a...	[(Basics, n), (norms, n), (set, v), (governmen...	Practical knowledge intake become less Onlin...
1	Online classes are beneficial but people belon...	Online classes are beneficial but people belon...	Online classes are beneficial but people belon...	Not at all	All stringency were taken aptly They all are w...	[(Online, n), (classes, n), (beneficial, a), (...	[(Online, n), (classes, n), (beneficial, a), (...	[]	[(stringency, n), (taken, v), (aptly, r), (wel...	Online class beneficial people belong lower ...

Fig 4.2: Dataframe Snapshot

4.4 Module 3: Proposed Algorithm to calculate Valence and Arousal

Input: Parsed Text Words

Output: Valence and Arousal vectors by proposed method

This module takes parsed text words as input and process them to find valence and arousal by the proposed method which uses SentiWordNet. For each topic-wise responses and combined responses the valence and arousal are calculated separately with the help of the following algorithm.

Proposed Algorithm for Valence and Arousal using SentiWordNet

1. Group collective textual responses of each participant: one per topic, and one combined.
2. For each topic, do the following:
    a. For each individual participant do the following:
        i. Initialize variables *var*, *valence* and *arousal* to 0.
        ii. Consider each word of the tokenized response and do the following:
            - Initialize variables *pos* and *neg* with positive and negative sentiment values of the word from SentiWordNet.
            - If the absolute difference of these values is more than 0 then
                - Increment *var* by 1.
                - Increment *valence* with (|*pos*| – |*neg*| +1)/2
                - Increment *arousal* with (|(|*pos*| – |*neg*|)|)
        iii. If *var* is greater than 0 then,
            - Divide *valence* by *var*
            - Divide *arousal* by *var*
        iv. Else store 0 in *valence* and *arousal* of the participant.
        v. Store value of *valence* to Valence vector.
        vi. Store value of *arousal* to Arousal vector.

4.5 Module 4: Dominance using BERT Summarizer and Cosine Similarity Function

Input: Raw Text from all Participants' Responses for Questions on all Survey Topics

Output: Dominance Value per Participant for each Topic in the form of Topic-wise Vectors

This module finds the dominance value from raw text by first applying extractive summarization techniques on the topic-wise agglomerative responses of all participants taken together and then comparing the similarity between this summary and the textual response of each participant individually as a cosine similarity value. The nearer an individual's response to the agglomerative summary, the higher the dominance factor for that individual.

Algorithm to find Dominance using Deep Machine Learning Techniques

1. Group agglomerative textual responses of all participants: one per topic, and one combined.
2. For each of these group topics ***g***, do the following:
    a. Use BERT Extractive Summarizer on the agglomerative response to generate a *group summarized response* of maximum 3 sentences.
    b. Use the pre-trained sentence-transformer model defined in python **'bert_base_nli_mean_tokens'** to generate the corresponding *group vector* ***Vg*** from the *group summarized response*.
    c. Now for each individual participant ***i***, do the following:
        i. Using the same sentence-transformer model find a vector ***Vig*** from the original collective textual response of the participant.
        ii. Find the Cosine Similarity between ***Vig*** and ***Vg.***
        iii. Store this value to space reserved on a Dominance Vector ***Dg***.
    d. Add the Dominance Vector to corresponding dataframe column

4.6 Module 5: VAD value extraction from NRC Lexicon (System 2)

In this scheme the valence, arousal and dominance of each participant are found by considering pre-determined scores of words using the NRC-VAD Lexicon defined by Saif Mohammad [10]. This Lexicon, which stores words in each line along with their corresponding Valence, Arousal and Dominance scores respectively in text format, needs a prolonged access time. This time is minimized by converting the lexicon into a dictionary at the outset.

4.6.1 Dictionary formation for NRC-Lexicon

- Read the NRC-Lexicon text file and split them based on new-line.
- Initialize three empty dictionaries which will store the valence, arousal and dominance score for the words.
- Traverse the NRC-Lexicon and get each word, along with its valence, arousal and dominance score in a list. The words and the scores can be distinguished by checking *tab* space.
- Save the valence of each word to the valence dictionary along with the word.
- Save the arousal of each word to the arousal dictionary along with the word.
- Save the dominance of each word to the dominance dictionary along with the word.

Now the valence, arousal and dominance score of all words present in NRC Lexicon has been imported to valence, arousal and dominance dictionary respectively.

4.6.2 Finding VAD using NRC Lexicon

Input: Parsed Text Words

Output: Valence, Arousal, Dominance vectors using NRC Lexicon

The algorithmic procedure to find VAD using NRC Lexicon:

1. Group collective textual responses of each participant: one per topic, and one combined.
2. For each topic, do the following:
    a. For each individual participant, do the following:
        i. Initialize variables *valence*, *arousal*, *dominance*, *var1*, *var2*, *var3* to 0.
        ii. For each word in the collective response for the participant, do:
            - If the valence score exists in valence dictionary, then add it to *valence* variable and increment *var1*.
            - If the arousal score exists in arousal dictionary, then add it to *arousal* variable and increment *var2*.
            - If the dominance score exists in dominance dictionary, then add it to *dominance* variable and increment *var3*.
        iii. Divide *valence* by *var1* and store the value to Valence vector.
        iv. Divide *arousal* by *var2* and store the value to Arousal vector.
        v. Divide *dominance* by *var3* and store the value to Dominance vector.

4.7 Module 6: Applying BERT Summarizer on Raw text of each topic

Input: Raw Text Responses of all participants on each topic

Output: Topic-wise Gist Vector for each participant

This module describes the first part of another proposed method which finds outliers from each participant's collective response text on a topic. Here one gist vector is found per topic for each participant by using BERT on the individual's collective topic-wise textual responses.

Algorithm to find Gist Vector from Raw Text for each participant using BERT

1. Group collective textual responses of each participant: one per topic, and one combined.
2. For each of these topical groups, do the following:
    a. For each individual participant, do the following:
        i. If the number of sentences in collective response is more than 3,
            Use summarizer to generate summary of the collective response;
            Else
            Consider the collective response as a whole.
        ii. Append the summarized/whole response to a list.

    b. Find Response Gist Vector of the appended list with the help of sentence transformer using predefined model 'bert-base-nli-mean-tokens'.
    c. Output each Response Gist Vector.

4.8 Module 7: Applying DBSCAN on raw text gist vectors

Input: Response Gist Vectors of all participants

Output: Cluster and Outliers

This module uses the gist response vectors obtained in the previous step and finds outliers with the help of cosine similarity and DBSCAN clustering.

DBSCAN algorithm for inter-participant similarity measures

1. Initialize epsilon distance, minimum number of elements as 2 to form a cluster.
2. For each participant's gist vector V.
    a. Pick each of the other gist vectors of responses V' at a time and do the following:
        i. Find Cosine Similarity between V and V'.
        ii. The distance or dissimilarity between the two response gist vectors is obtained by subtracting the similarity from 1:
        Distance = 1 – Cosine Similarity
3. Perform DBSCAN algorithm on the obtained distance values and generate cluster and outliers amongst the participants.

4.9 Module 8: DBSCAN on VAD vectors

Input: Topic-wise VAD vectors for all participants

Output: Topic-wise Outliers

This module finds outliers based on the input VAD vectors.

DBSCAN with V-A-D vectors

1. For each topic do the following:
    a. Consider the values of valence, arousal and dominance scores of all participants
    b. Import DBSCAN from scikit-learn package.
    c. Fit the valence, arousal and dominance values along with minimum number of points to form a cluster and epsilon distance to the function to get the outliers.

4.10 Module 9: Finding Accuracy of proposed System 1 using VAD values

Input: 1) V,A,D vectors for different topics from proposed System 1

2) V,A,D vectors for different topics from proposed System 2

Output: Accuracy Tables and Bar charts

This module tries to find out the accuracy of the proposed model by comparing the VAD vector obtained from the proposed model with VAD vector obtained by using NRC Lexicon.

V-A-D Vectors compared using similarity measures

1. For each topic, do the following:
    a. For valence, arousal and dominance each at a time, do the following:
        i. Get the corresponding vectors of all participants obtained from the two methods (System 1 and 2).
        ii. Find cosine similarity between these two vectors.
        iii. The obtained similarity value is placed appropriately within a table.
    b. Topic-wise Accuracy tables gets displayed
2. Collective Accuracy tables and bar charts gets displayed

4.11 Module 10: Comparison and Combination of Outliers

Input: Outliers obtained from System 1, System 2 and System 3

Output: Outlier Comparison through Tables and Venn Diagrams

This module compares the outliers found by all the three methods. The obtained outliers are compared topic-wise to find the common outlier(s) detected by all three systems. This helps to understand the performance of the proposed models collectively. The results are visualized with the help of Tables and Venn-diagrams.

The next chapter records the overall system performance analytically.

Chapter 5: Results and Performance Analysis:

This chapter contains the description of the data and tools used to generate the proposed systems. It also contains a detailed discussion on the outputs obtained from those systems, followed by a comprehensive performance analysis made with graphical representations.

5.1 Data Description

The data is obtained from a survey that has been conducted among the university students. They were asked to answer 12 different questions related to Covid-19. The survey was conducted online with the help of Google Form, from which the data is extracted in csv format. There were 77 participants in all. The questions were from three different topics – from problems related to education, economy and general aspects faced by society during Covid-19.

5.2 Tools Utilised

5.2.1 Hardware Requirements

- System: HP Laptop
- Processor: Dual Core Intel Core i5 4210U @ 1.70Ghz
- RAM: 8.00 GB
- System type: Windows 10, 64-bit Operating System

5.2.2 Software Requirements

- Python of version greater than 3.6 or higher
- Integrated Development Environments like Jupyter Notebook, where Python Programming is done

5.3 Results and Inferences

The outputs derived from the three systems studied in this work are presented in this section. The performance analysis of the proposed BERT based techniques (Systems 1 and 3) are made – with benchmarking against the standard NRC Lexicon-based system 2.

5.3.1 Outputs from System 1

As already discussed in the Methodology Chapter 4, System 1 uses a method proposed to find Valence and Arousal values of student responses from SentiWordNet positive-negative sentiment values for constituent text words. In this system, the Dominance value is extracted directly from raw text by applying BERT techniques. The observations on different topics from System 1 are described below. The topics include those related to education, economy, general aspects, as well as a combination of all these.

A. Results for Combined topic

A sample of the valence, arousal and dominance values from the combined responses for some of the participants are displayed below in figure 5.1. These are depicted in the form of a screenshot containing a three-column table, in which the columns represent valence, arousal and dominance of each person's response respectively. The values are all fractional numbers lying in the range 0 to 1.

	A	B	C	D
1	Valence	Arousal	Dominance	
2	0.562500	0.352273	0.877000	
3	0.604167	0.475000	0.889264	
4	0.565645	0.329403	0.827451	
5	0.559932	0.380137	0.840341	
6	0.484375	0.368750	0.808090	
7	0.503425	0.363014	0.835110	
8	0.605263	0.342105	0.825892	
9	0.570175	0.412281	0.835777	
10	0.542553	0.329787	0.777712	
11	0.525310	0.364669	0.838268	
12	0.361111	0.333333	0.546639	
13	0.611111	0.305556	0.754729	
14	0.528125	0.327083	0.827920	
15	0.535000	0.310000	0.883878	
16	0.549107	0.431548	0.862527	
17	0.553711	0.326172	0.855704	
18	0.642857	0.367347	0.811630	
19	0.574695	0.356707	0.798743	
20	0.518429	0.389423	0.826444	
21	0.532227	0.380859	0.805000	
22	0.562500	0.346429	0.694610	
23	0.495427	0.350610	0.869044	

Fig 5.1: Sample Screenshot of VAD values obtained by System 1

The complete set of outputs have been used to detect outlier personalities directly, as shown in figure 5.2 below – with another screenshot representing the outlier participants as numbers ranging between 1 and 77. The algorithm used here is DBSCAN as has been already explained in section 4.9 of the previous chapter. Here the algorithm sets minimum number of points to form a cluster to 2 and epsilon distance to 0.08.

```
[11, 41, 43, 46]
```

Fig 5.2: Outliers obtained from System 1 on Combined topic

B. Results for Education topic

A similar depiction is presented below in figure 5.3, consisting of outliers obtained with responses based on education. Here, minimum number of points to form a cluster remains the same i.e., 2 and epsilon distance is set to 0.10.

Fig 5.3: Outliers obtained from System 1 on Education topic

C. Results for Economic topic

The responses based on economic topic are processed by system 1 with minimum number of points remaining the same as before and epsilon distance set to 0.2. The result obtained is described below in figure 5.4.

Fig 5.4: Outliers obtained from System 1 on Economic topic

D. Results for General topic

The responses based on general topic are processed by system 1, resulting in the following outliers shown in figure 5.5 below. Here the epsilon distance is reset to 0.10.

Fig 5.5: Outliers obtained from System 1 on General topic

5.3.2 Outputs from System 2

The System 2 uses NRC Lexicon to find valence, arousal and dominance from responses within the same dataset as per the module already described in section 4.6 of the previous chapter. Using these values, outliers are again found applying DBSCAN technique with minimum number of points to form a cluster set to 2.

A. Results for Combined topic

The VAD values obtained using NRC Lexicon are used to find outliers on combined topic by setting epsilon distance as 0.036, as described below in figure 5.6.

Fig 5.6: Outliers obtained from System 2 on Combined topic

B. Results for Education topic

The VAD values obtained using NRC Lexicon are used to find outliers on educational topic by setting epsilon distance as 0.041, as described below in figure 5.7.

```
[11, 12, 41, 48, 70]
```

Fig 5.7: Outliers obtained from System 2 on Education topic

C. Results for Economic topic

The VAD values obtained using NRC Lexicon are used to find outliers on economic topic by setting epsilon distance as 0.04, as described below in figure 5.8.

```
[4, 11, 25, 39]
```

Fig 5.8: Outliers obtained from System 2 on Economic topic

D. Results for General topic

The VAD values obtained using NRC Lexicon are used to find outliers on economic topic by setting epsilon distance as 0.041, as described below in figure 5.9.

```
[11, 12, 33, 43, 46, 52]
```

Fig 5.9: Outliers obtained from System 2 on General topic

5.3.3 Comparative Accuracy: System 1 vs System 2

This section involves the performance test of the proposed method for VAD. For each response, two sets of valence, arousal and dominance values have been found using two methods, System 1 applying the proposed method using SentiWordNet and BERT, and System 2 which uses NRC Lexicon. These values are compared to evaluate the accuracy of the proposed model described in System 1. The better the similarity among VAD vector sets, the better the accuracy of the proposed model.

A. VAD Accuracy table for Combined topic

The similarity of VAD values obtained from the two methods on Combined topic is displayed in the table 5.1 below:

Table 5.1: VAD Accuracy table based on Combined responses

Parameter	**Accuracy**
Valence	99.6765 %
Arousal	98.7989 %
Dominance	99.4885 %

Inference: In case of combined responses to individual topics, the obtained similarity between the proposed method and existing dictionary-based technique is found to be quite high – approximately 99 %.

B. VAD Accuracy table for Education topic

The educational responses are considered next. Here the VAD values of both the methods are compared on educational topic. The result is displayed below in table 5.2:

Table 5.2: VAD Accuracy table based on Education responses

Parameter	**Accuracy**
Valence	99.3184 %
Arousal	97.5286 %
Dominance	99.1476 %

Inference: For responses to Education related questions, the obtained similarity score of valence, arousal and dominance are also quite high. This indicates the accuracy of the proposed model is maintained in educational responses also.

C. VAD Accuracy table for Economy topic

The similarity between the two methods on Economic responses are described in the table 5.3 given below:

Table 5.3: VAD Accuracy table based on Economical responses

Parameter	**Accuracy**
Valence	79.6020 %
Arousal	75.2276 %
Dominance	92.3278 %

Inference: From the above table, it is observed that valence and arousal for economy topic do not have as good a similarity as dominance. This may be because, in the survey there was only one question related to economy. So, data is far less for this topic and hence the accuracy on economic response for valence and arousal is unable to reach a satisfactory point.

D. VAD Accuracy table for General topic

Similarity between the two methods on General responses are described in the table 5.4 below:

Table 5.4: VAD Accuracy table based on General responses

Parameter	**Accuracy**
Valence	98.5924 %
Arousal	96.5803 %
Dominance	97.4097 %

Inference: From the above similarity measures, it is observed that on general responses too, the proposed technique gives satisfactory results.

E. Collective result

The accuracies obtained are collectively presented in table 5.5 below.

Table 5.5: Collective VAD Accuracy values from System 1 and System 2

Parameter	Accuracy			
	Combined	Educational	Economic	General
Valence	99.9765 %	99.3184 %	79.6020 %	98.5924 %
Arousal	98.7989 %	97.5286 %	75.2276 %	96.5803 %
Dominance	99.4885 %	99.1476 %	92.3278 %	97.4097 %

The result has also been summarized with the help of a bar chart depicted in figure 5.10 below.

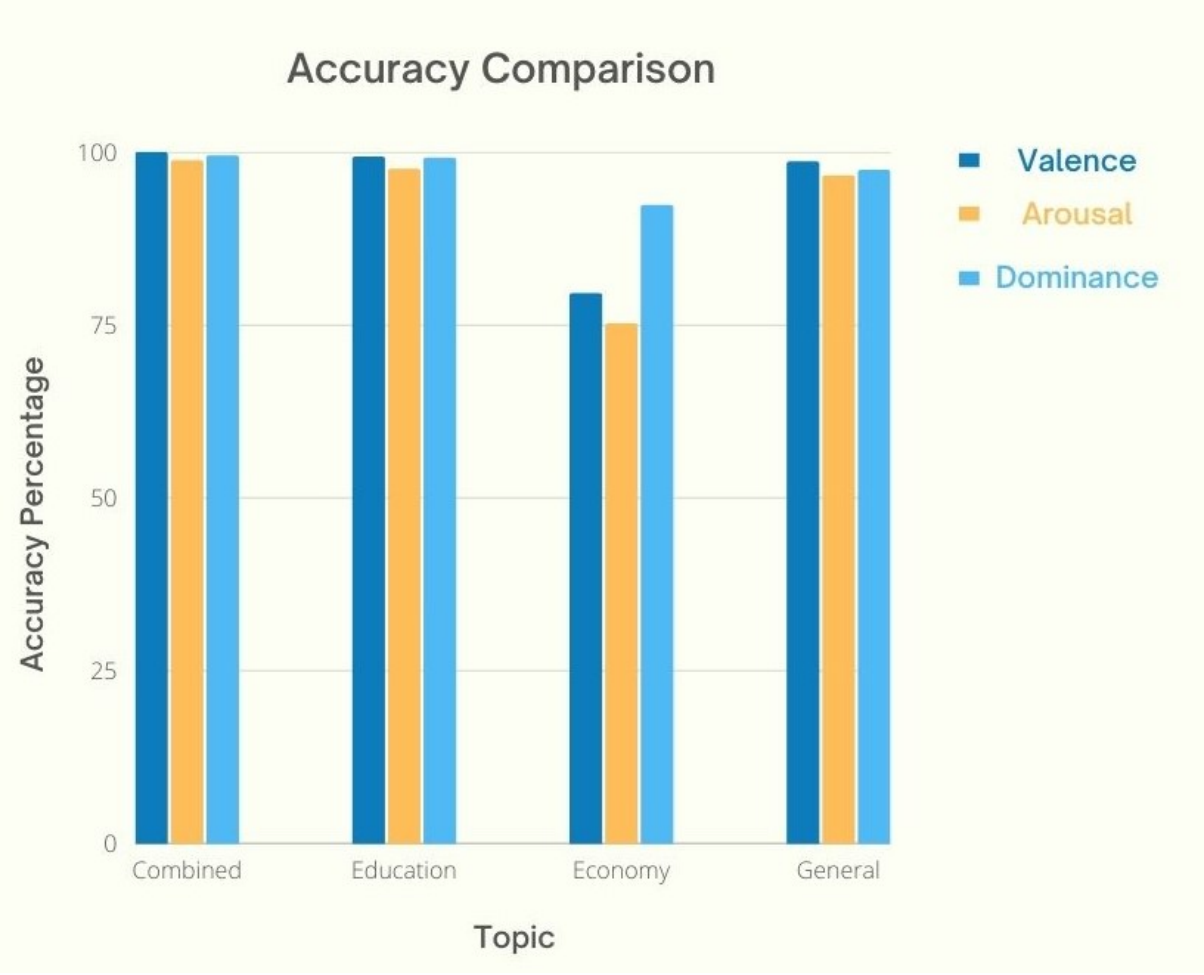

Fig 5.10: Graphical Comparison of VAD Accuracy between Systems 1 and 2

General Inference: As already inferred, the overall accuracy on the combined set is the best as it considers the whole dataset, and economy fairs poorly due to scarcity of data. Hence, among the four different topics, combined, educational, economic and general, the highest similarity between the two systems is found in combined responses and the least is obtained in the economic topic.

5.3.4 Output of System 3

The proposed System 3 produces sets of outliers on different topics using BERT on raw text and considering semantic content, the output of which will be discussed next. Here, DBSCAN technique is used as already discussed in section 4.8 of previous Chapter 4, with minimum number of points to form a cluster being set to 2 for all topics.

The outliers produced by this system on the different topics, as well as the combined one, are displayed in table 5.6 below.

Table 5.6: Outliers obtained on different topics using System 3

Topic	**Obtained Outliers**
Combined	11, 13, 25, 41
Education	11, 41, 43, 48, 58, 63
Economy	21, 49, 62, 67
General	41, 48, 56, 67

In each case, the remaining participants are mapped to some cluster or other.

5.3.5 Final Output

In this section, the outliers obtained from the three different systems are being compared topic-wise.

A. Outliers for Combined topic

The outliers obtained from combined responses by all three systems are displayed in the following table 5.7. Outliers common to all three systems are highlighted with colour red and the outliers common within systems 1 and 2 only are coloured in orange.

Table 5.7: Outliers obtained on Combined responses from all 3 Systems

Used Method	**Obtained Outliers**
System 1	11, 41, 43, 46
System 2	11, 12, 41, 46, 53, 56
System 3	11, 13, 25, 41

The output of the above table is represented pictorially as a Venn-diagram in the following figure 5.11. This helps to highlight the overlapped outliers across the system boundaries in a precise manner.

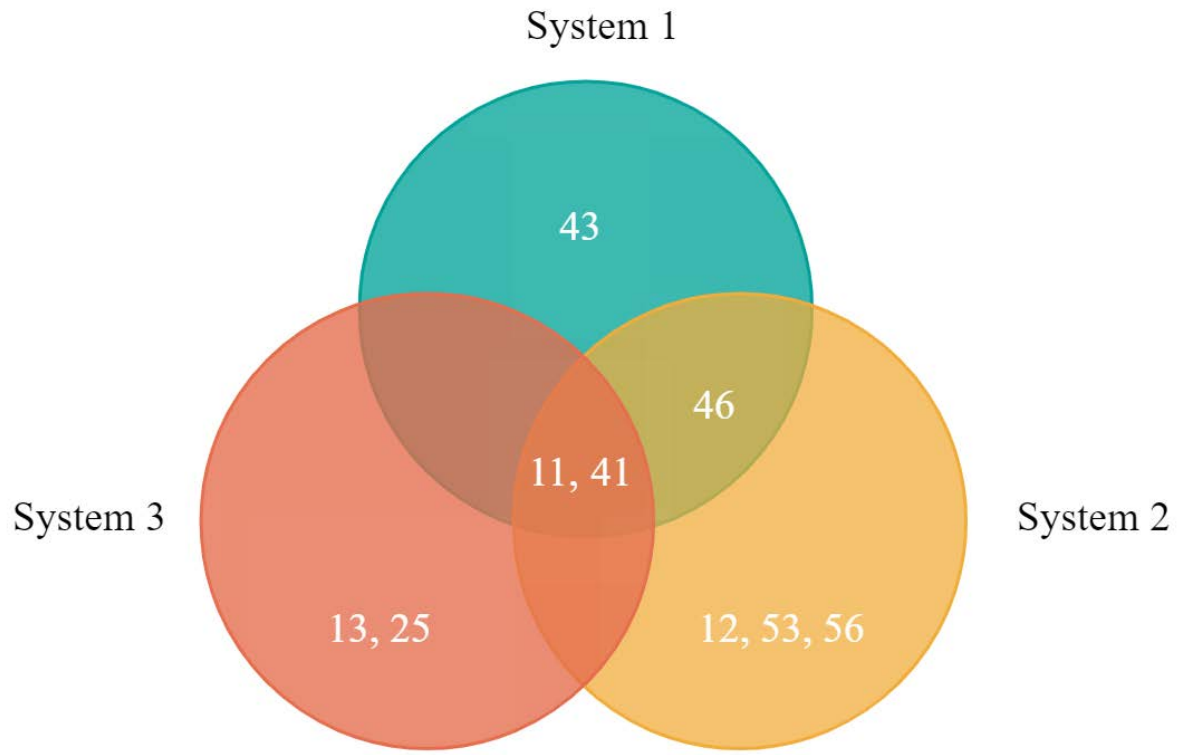

Fig 5.11: Venn diagram on the Combined topic

From the above table and diagram, it is observed that students 11 and 41 are identified as outliers by all three systems on combined topic. The original and summarized responses of these students are given below in tables 5.8 and 5.9.

Table 5.8: Original and Summarized Response of Student 11 on Combined topic

Original Response	Summarized Response
early exposure to technology. ,casual attendance. ,it was need of the hour. ,it's useless. ,no. ,neither benefit nor jolt. ,lockdown.,covid norms are maintained. ,it was needed. ,horrible. ,yes. ,yes. ,early exposure to technology. casual attendance. it was need of the hour. it's useless. no. it was needed. ,neither benefit nor jolt. ,lockdown. covid norms are maintained. horrible.	early exposure to technology. ,it's useless. ,no. ,neither benefit nor jolt. ,covid norms are maintained. ,it was needed. ,horrible. it's useless. no. it was needed. ,neither benefit nor jolt. covid norms are maintained. horrible.

Inference: Both the original and summarized response of student 11 contain many negative words and phrases like *useless, horrible, neither benefit nor jolt*, which may have caused the response to be an outlier. The response as a whole indicates negativity.

Table 5.9: Original and Summarized Response of Student 41 on Combined topic

Original Response	Summarized Response
connectivity issues,comfortable,nope,wearing mask and maintaining social distancing,no comment,terrible,ofcourse yes,definitely,connectivity issues. comfortable. nan. nan. nope. no comment,nan,nan. wearing mask and maintaining social distancing. terrible. ofcourse yes. definitely	connectivity issues,comfortable,nope,wearing mask and maintaining social distancing,no comment,terrible,ofcourse yes,definitely,connectivity issues. wearing mask and maintaining social distancing.

Inference: Use of words like *issues* and *terrible* may have marked student number 41 as an outlier. Here too both the original and summarized responses convey negative impression.

The following sections explores the individual topic-wise outliers.

B. Outliers for Educational topic

The following table 5.10 describes the outliers obtained from education topic by all three systems.

Table 5.10: Outliers obtained on Educational Responses from all 3 Systems

Used Method	**Obtained Outliers**
System 1	11, 12, 41, 46, 48, 56
System 2	11, 12, 41, 48, 70
System 3	11, 41, 43, 48, 58, 63

The following Venn-diagram in figure 5.12 depicts the outlier overlaps clearly again for the Education topic. This topic considers the largest number of questions, and as such produces the best results amongst the individual topics in the form of three common outliers marked in red in the table 5.10 above.

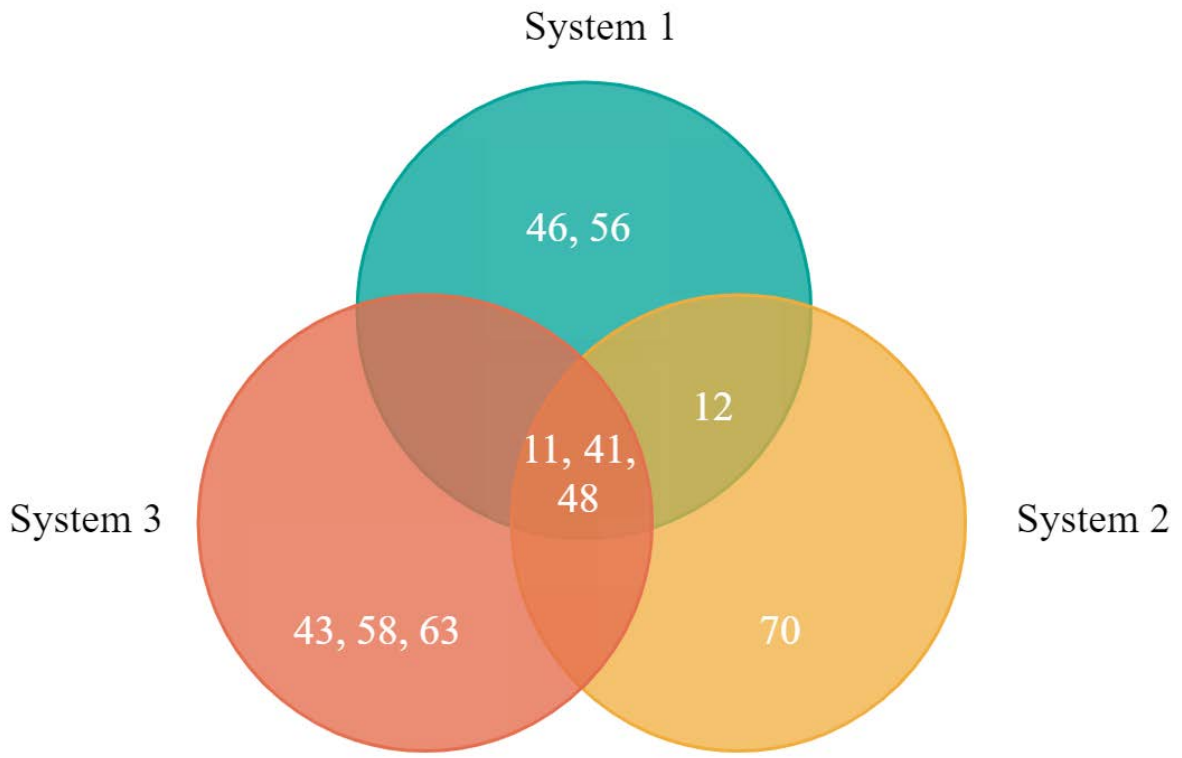

Fig 5.12: Venn diagram on Education topic

Responses of the three students 11, 41 and 48, identified as outliers by all three systems on the educational topic, are described below in tables 5.11, 5.12 and 5.13.

Table 5.11: Original and Summarized Response of Student 11 on Education topic

Original Response	Summarized Response
early exposure to technology. casual attendance. it was need of the hour. it's useless. no. it was needed.	early exposure to technology. it was need of the hour. it's useless. no. it was needed.

Inference: Negative words such as *no* and *useless* may have marked the student to be an outlier.

Table 5.12: Original and Summarized Response of Student 41 on Education topic

Original Response	Summarized Response
connectivity issues. comfortable. nan. nan. nope. no comment	connectivity issues. comfortable. nan. nan. nope. no comment

Inference: Again, usage of words like *issues*, *no*, *nope* may have caused this to be an outlier.

Table 5.13: Original and Summarized Response of Student 48 on Education topic

Original Response	Summarized Response
online class is best . covid has many detrimental effects . through online we can revise a topic many times. practical also in online form. yes. i think reopen is very much risky	covid has many detrimental effects . i think reopen is very much risky

Inference: Words such as *detrimental* and *risky* may have marked student 48 as an outlier.

C. Outliers for Economic topic

Following table 5.14 describes the outliers obtained from economic topic by all three systems.

Table 5.14: Outliers obtained on Economic responses from all 3 Systems

Used Method	**Obtained Outliers**
System 1	11, 56, 63
System 2	4, 11, 25, 39
System 3	21, 49, 62, 67

Following Venn-diagram in Fig. 5.13 depicts the overlap between outliers for Economy.

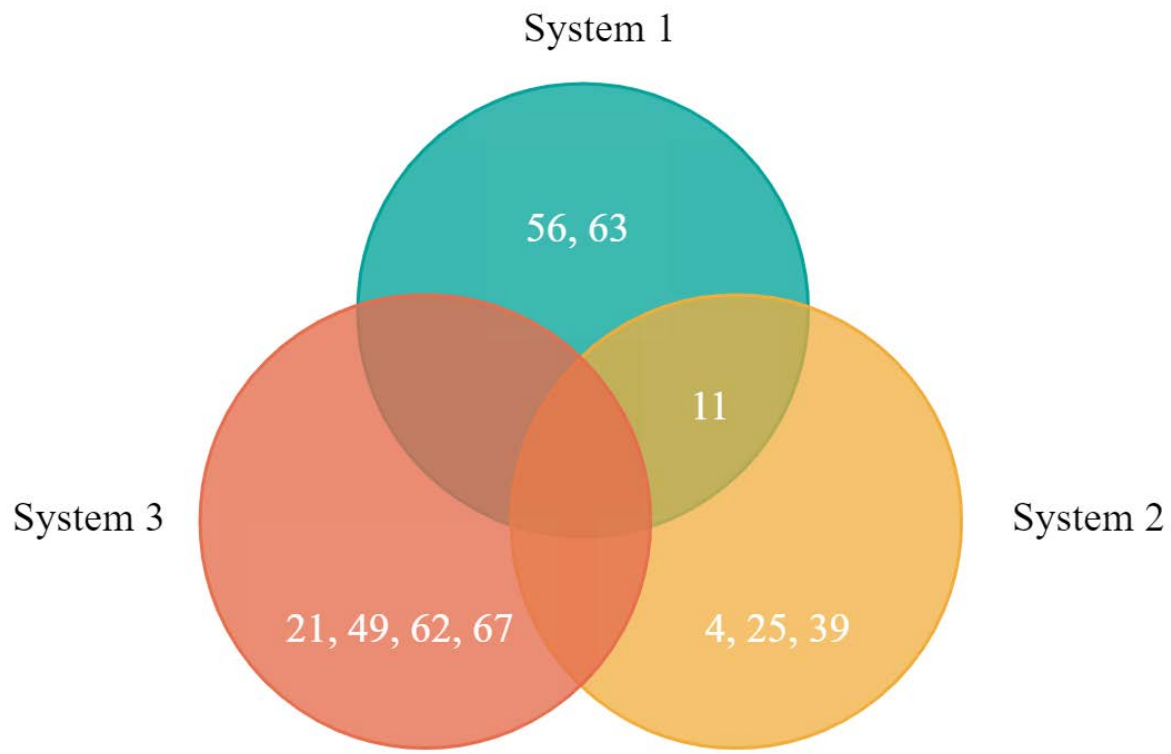

Fig 5.13: Venn diagram on Economic topic

From the above table and diagram, it is found that there is no common outlier found by the three systems on economy topic. This may be due to dearth of data. For this topic, none of the system 3 outliers could be matched with the other two systems. Systems 1 and 2 detect one common outlier in student 11.

D. Outliers for General topic

The following table 5.15 describes the outliers obtained from general topic by all three systems. Outliers common to systems 2 and 3 are marked in blue.

Table 5.15: Outliers obtained on General responses from all 3 Systems

Used Method	Obtained Outliers
System 1	11, 12, 32, 43, 46
System 2	11, 12, 33, 43, 46, 52
System 3	43, 52, 56, 67

The Venn-diagram in Fig. 5.14 represents the overlapped outliers on general topic texts.

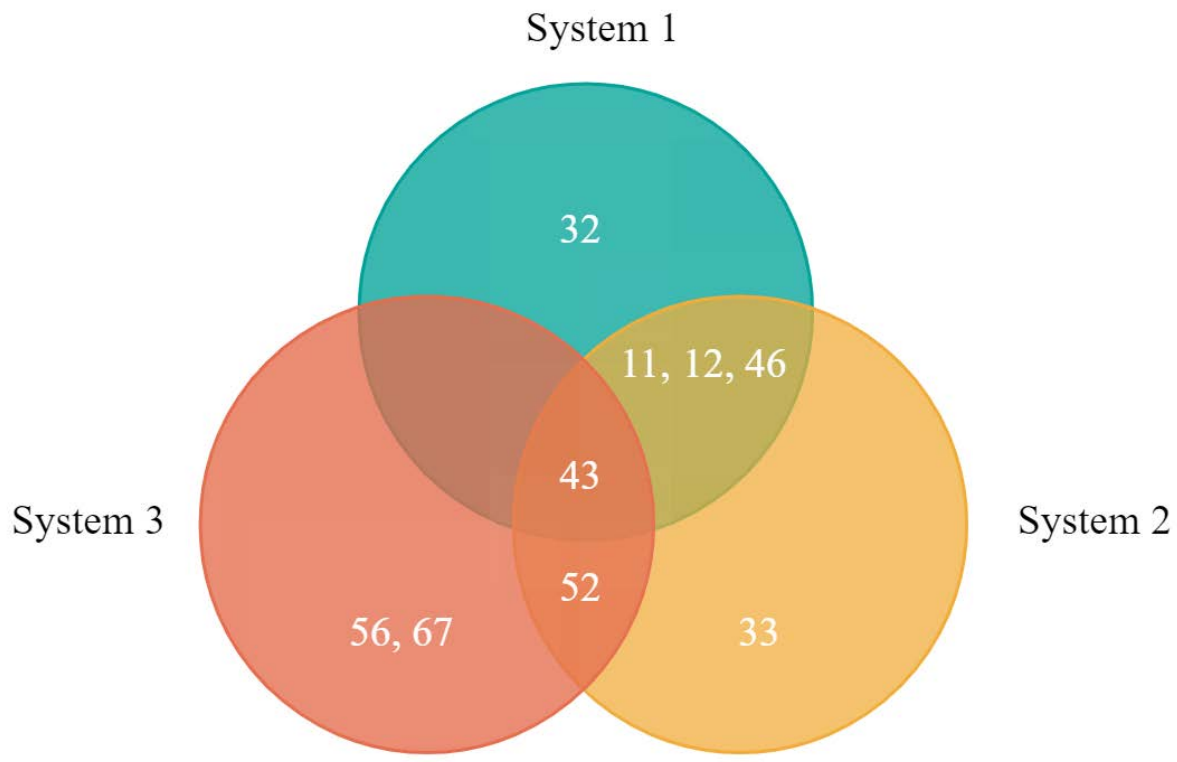

Fig 5.14: Venn diagram on General topic

From the above table and diagram, it is found that student 43 is marked as outlier by all three systems for general topic. The original and summarized response of the same is given below table 5.16.

Table 5.16: Original and Summarized Response of Student 43 on General topic

Original Response	Summarized Response
government imposed stringency is right in way. good. it's not good experience . game changer. yes strict	government imposed stringency is right in way. it's not good experience . game changer.

Inference: Use of words and phrases like *stringency*, *not good* has made the response an outlier.

5.3.6 Overall Topic-wise Outlier Alerts

Outliers obtained on different topics by all the three systems are described below in table 5.17, along with the coloured alerts. Participant numbers marked in red are found to be in highest level of alert, participant numbers marked in orange and blue are found to be in medium level of alert and participant numbers marked in black are found to be at the lowest level of alert. Highest level of alert indicates immediate attention.

Table 5.17: Overall Outlier Alerts found from All topics

Topic	**Outliers**
Combined	11, 12, 13, 25, 41, 43, 46, 53, 56
Education	11, 12, 41, 43, 46, 48, 56, 58, 63, 70
Economy	4, 11, 21, 25, 39, 49, 56, 62, 63, 67
General	11, 12, 32, 33, 43, 46, 52, 56, 67

Inference: From the above table 5.17, it is observed that student 11 has been marked as outlier by all the systems. For combined and education topic, the student is marked with high alert, and for economy and general topic with medium alert. This indicates that student 11 should be provided immediate attention to avoid any unpleasant outcome in near future. Additionally, students 41, 43 and 48 also seeks attention as candidates with high alert. Students 12, 46 and 52 have medium alerts associated with them. The rest of the outliers may also be considered although the level of attention is indicated to be low in their cases.

The next chapter concisely presents the overall goal achieved through this research. It also points towards some new tools and techniques that may be employed to improve performance measures.

Chapter 6: Conclusion and Future Scope

At the very introductory chapter it was hinted that this work aims to address a crisis, threatening today's mankind in general – and the student community in particular – which is the enemy within one's own self. In short, it is the failure to combat internal emotions that is driving humans crazy in this tough modern world. Like physical ailments, such mental ailments also eat into the core of a person's existence. But often, earlier diagnosis helps to heal up the patches more efficiently and effectively. So, the goal of this thesis is specifying pervasive techniques to assess the degree of mental disorientation in advance.

The proposed work utilises three different methods to find outlier personalities by analysing textual survey entries of university students in response to COVID-19 related questions. All three systems have been tested with DBSCAN technique for finding outlier personalities. The first two systems detect the same by considering VAD values, while the third segregates them directly using BERT on raw inputs. In the very first system, Valence and Arousal are directly calculated on the basis of a proposed statistical measure using positive-negative emotion values of words from a standard lexicon. Here Dominance is calculated using BERT and Cosine Similarity functions. The second system, purely based on annotated lexicons, is presented to benchmark the first proposed system. The third one is another proposed system which discovers outliers from raw texts, rather than from VAD values. The trace of commonality presented by the three systems in their outputs is captured by visual aids such as Venn-diagrams. Alerts come in the form of colour coded entries in tables. So, it is felt that the objective of finding out alienated students have been successfully met to a large degree.

Furthermore, the overall similarity between VAD values obtained from systems 1 and 2 have been tested and found to be satisfactory. The best comparison results for valence, arousal and dominance are 99.97%, 98.79% and 99.48% respectively. These are obtained from the Combined and the Education topic, both of which understandably win the race by dint of possessing bulkier datasets. The Economy factor, on the other hand, scores the lowest just because of its data dearth – only one question being allotted to this topic, and that too avoided by most of the participants. These revelations have also been further corroborated by the data visualization techniques affected through performance bars and Venn-diagrams presented in the result section of the previous chapter.

It is felt that the outliers found by the three systems may be improved further by utilising a larger dataset and by trying out various types of summarizers including abstractive ones. Extracting finer-grained emotions within text may also help to improve the outlier detection process. Audio responses can be considered in addition to textual responses in this context. Dictionaries other than SentiWordNet and NRC lexicon can be used to reevaluate the performance of the proposed techniques. Additionally, a plethora of Deep Learning techniques can also be explored in this connection by studying the enriched survey paper on the topic discussed by D.W. Otter et al [38].

References

1. Xing Fang, Justin Zhan, “Sentiment analysis using product review data”, Journal of Big Data - Springer Open Journal, 2015
2. Federico Neri, Carlo Aliprandi, Federico Capeci, Montserrat Cuadros, “Sentiment Analysis on Social Media”, IEEE/ACM International conference on advances in social networks analysis and mining, 2012
3. Lei Wang, Jianwei Niu, Shui Yu, “SentiDiff: Combining textual information and sentiment diffusion patterns for twitter sentiment analysis”, IEEE Transactions on knowledge and data engineering, 2018
4. Jacob Devlin, Ming-wei Chang, Kenton Lee, Kristina Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding”, arXiv, 2018
5. Osgood C., Suci G., and Tannenbaum P.,“The measurement of meaning”. University of Illinois Press, 1957
6. Russell J., “A circumplex model of affect”, Journal of personality and social psychology, Vol. 39, No. 6, 1161-1178, 1980
7. Russell J., “Core affect and the psychological construction of emotion”, Psychological Review, Vol. 110, No. 1, 145-172, 2003
8. Bradley M, and Lang P., “Affective norms for English words (ANEW): Instruction manual and affective ratings”, University of Florida, 1999
9. Warriner A., Kuperman V. and Brysbaert M., “Norms of valence, arousal, and dominance for 13,915 English lemmas”, 2013
10. Mohammad S., “Obtaining Reliable Human Rating of Valance, Arousal and Dominance for 20000 english word”, National Research Council Canada, 2018
11. Chuhan W., Fangzhao W., Sixing W., Zhigang Y., Junxin L., and Huang Y. “Semisupervised dimensional sentiment analysis with variational autoencoder”, Knowledge-Based Systems, Vol. 165, 30-39, 2019
12. Suyang Z., Shoushan L., and Guodong Z., “Adversarial attention modelling for multidimensional emotion regression”, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 471-480, 2019
13. Akhtar S., Ghosal D., Ekbal A., Bhattacharyya P., and Kurohashi S., “All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework”, 2019
14. Moors A., Houwer J., Hermans D., Wanmaker S., Schie K., Harmelen A., Schryver M., Winne J., and Brysbaert M., “Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words”, Behav Res 45, 169-177, 2013
15. Melissa L., Conrad M., Kuchinke L., Urton K., Hofmann M., and Jacobs A., “The berlin affective word list reloaded”, Behavior Research Methods 41, 534-538, 2009
16. Redondo J., Fraga I., Padron I, and Comesana M., “The spanish adaptation of anew (affective norms for English words)”, Behavior Research Methods 39, 600-605, 2007

17. Andrea Esuli, Fabrizio Sebastiani, “SentiWordNet: A publicly available lexical resource for Opinion Mining”, Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06), European Language Resources Association (ELRA), 2006.
18. Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani, “SentiWordNet 3.0: An enhanced lexical respirce for sentiment analysis and opinion mining”, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10), European Language Resources Association (ELRA), 2010.
19. Shailendra Singh, Sanchita Paul, “Sentiment Analysis of Social Issues and Sentiment Score Calculation of Negative Prefixes”, International Journal of Applied Engineering Research, vol 10, 1694-1699, 2015
20. H.P. Luhn, “The Automatic Creation of Literature Abstracts”, Presented at IRE National Convention, New York, 159-165, 1958
21. Joel larocca Neto, Alex A. Freitas and Celso A.A.Kaestner, "Automatic Text Summarization using a Machine Learning Approach”, Book: Advances in Artificial Intelligence: Lecture Notes in computer science, Springer Berlin / Heidelberg, Vol 2507/2002, 205-215, 2002.
22. J. Ross Quinlan. "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, Inc., 1993
23. Barzilay R., & McKeown K. R. “Sentence Fusion for Multidocument News Summarization”. Computational Linguistics, 31(3), 297–328, 2005
24. R. Khan, “Extractive based Text Summarization Using K-Means and TF-ID”, I.J. Information Engineering and Electronic Business, vol 3, 33-44, 2019
25. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, “RoBERTa: A robustly optimized BERT Pretraining Approach”, ICLR 2020
26. Alfirna Rizqi Lahitani, Adhistya Erna Permanasari, Noor Akhmad Setiawan, “Cosine Similarity to Determine Similarity Measure: Study Case in Online Essay Assesment”, 4th International Conference on Cyber and IT Service Management, 2016
27. Martin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, KDD’96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 226-231, 1996
28. Dingsheng Deng, “DBSCAN Clustering Algorithm Based on Density”, 7th International Forum on Electrical Engineering and Automation (IFEEA), 2020
29. Tom M. Mitchell. Book: Machine Learning
30. Arthur Lee Samuel, “Some Studies in Machine Learning Using the Game of Checkers”, IBM Journal of Research and Development – vol 3: 210-229
31. Warren McCulloch, Walter Pitts, “A Logical Calculus of Ideas Immanent in Nervous Activity”, Bulletin of Mathematical Biophysics 5:115-133, 1943

32. Osisanwo F.Y., Akinsola J.E.T., Awodele O., Hinmikaiye J.O., Olakanmi O., Akinjobi J., “Supervised Machine Learning Algorithms: Classification and Comparison”, International Journal of Computer Trends and Technology (IJCTT) – vol 48, 2017
33. Prakash M Nadkarni, Lucila Ohno-Machado, Wendy W Chapman, “Natural Language Processing: An Introduction”, Journal of the American Medical Informatics Association – vol 18: 544-551, 2011
34. Mehrabian, A., & Russell J., “Evidence for a Three-Factor Theory of Emotions”, Journal of Research in Personality, 273-294, 1977
35. https://iq.opengenus.org/
36. https://www.analyticsvidhya.com/
37. https://www.nltk.org/
38. Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita, “A Survey of the Usages of Deep Learning for Natural Language Processing”, IEEE Transactions on Neural Networks and Learning Systems, 2020