Hate Speech and Negative Emotions Detection System for Social Media Comments Using BERT and BiLSTM Models hyperparemeter tuned with Particle Swarm Optimization

Submitted in partial fulfilment of the requirements of the degree of
Bachelor of Technology (B.Tech)

by

Mohd Sufiyan Ansari (197151)
Maram Mahitha (197249)
Ahana Panja (197206)

Under the supervision of

Dr. T. Ramakrishnudu
Associate Professor, CSE Department

Department of Computer Science and Engineering

National Institute of Technology

Warangal

2022-23

APPROVAL SHEET

This Dissertation Work entitled **"Hate Speech and Negative Emotions Detection System for Social Media Comments Using BERT and BiLSTM Models hypertuned with Particle Swarm Optimization"** by
Mohd Sufiyan Ansari (197151)
Maram Mahitha (197249)
Ahana Panja (197206)
is approved for

the degree of Bachelor of Technology (B.Tech).

Examiners

____

____

____

Supervisor

____

Dr. T. Ramakrishnudu

Chairman

____

Dr. S. Ravi Chandra (HOD, CSE Dept.)

Date : ____________________

Place : ____________________

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY

WARANGAL

CERTIFICATE

This is to certify that the Dissertation work entitled "**Hate Speech and Negative Emotions Detection System for Social Media Comments Using BERT and BiLSTM Models hypertuned with Particle Swarm Optimization**" is a bonafide record of work carried out by **Mohd Sufiyan Ansari (197151)**, **Maram Mahitha (197249)** and **Ahana Panja (197206)**, submitted to the Dr. T. Ramakrishnudu of **Department of Computer Science and Engineering**, in partial fulfilment of the requirements for the award of the degree of B. Tech at **National Institute of Technology, Warangal** during the 2022-2023.

Dr. T. Ramakrishnudu	Dr. S. Ravi Chandra
Project Guide	Head of the Department
Department of CSE	Department of CSE
NIT Warangal	NIT Warangal

Acknowledgement

We would like to express our heartily gratitude towards **Dr. T. Ramakrishnudu**, Associate Professor, CSE Department, for his valuable guidance, supervision, suggestions, encouragement and the help throughout the semester and also for the completion of our project work. He kept us going when we were down and gave us the courage to keep moving forward.

We would like to take this opportunity once again to thank **Dr. S. Ravi Chandra**, Head of the Department, Computer Science and Engineering, NIT Warangal for giving us this opportunity and resources to work on this project and supporting through out. We also want to thank evaluation committee for their valuable suggestions on my proposals and research and for conducting smooth presentation of the project.

Mohd Sufiyan Ansari	Maram Mahitha	Ahana Panja
197151	197249	197206

Declaration

We declare that this written submission represents our ideas, our supervisor's ideas in our own words and where others' ideas or words have been included, We have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Mohd Sufiyan Ansari	Maram Mahitha	Ahana Panja
197151	197249	197206

Abstract

Social Media has been a great way for people to stay connected amidst their busy lives. The audience includes people from all age groups - children, teenagers, adult. People share their views publicly on such platforms which may or may not affect the mental condition of other person reading the same.

Texts/paragraphs containing words which are abusive or threatening or expresses prejudice on the basis of ethnicity, religion, gender, or similar grounds referred to as Hate Speeches. The increase in presence of Hate Speech on social media has negative influence on the masses and has also been linked to rise in global violence.

The primary goal of this project is develop a model to identify and detect the presence of hate speech/quotation in the comments and posts that are made on public platforms with greater efficieny and accuracy. The next objective is to develop a flagging system that will use the proposed model to detect such objectionable text real-time and generate an alert message for the user before the text is posted on social media platforms.

The proposed model uses Bi-directional Encoder Representation for Transformers (BERT) with a Bidirectional Long Short Term Memory (Bi-LSTM) model is built on top for accomplishing the use case mentioned above.

An issue with training models such as our BERT+BiLSTM is searching and selection of the most optimal set of hyperparameters.A suitable Hyperparameter Optimization Algorithm is required because for performing the same task on the same datasets different choices of hyperparameters produce wide variance in test results.

The Particle Swarm Optimization PSO algorithm is well suited for tuning complex functions and converges to global optimum in less number of iterations. The PSO algorithm is used in the proposed model for fine-tuning BERT.

The accuracy obtained from running the BERTBiLSTM model with PSO hyperparameter tuning(93.1%) is better than just the BERTBi-Lstm model accuracy(83%).

Keywords —- Sentiment Analysis, Flagging System, BERT, NLP, Bi-LSTM, Hyperparameter Optimization, Metaheuristics, Particle Swarm Optimization

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Problem

1.1.1 Social Media

Social media can be defined as the online interactive platforms available to the public all over the world. This helps people to connect with each other on a global level[1]. People can exchange their ideas and opinions and gain more knowledge. Selling and marketing products across the world is being made possible by social media. Social media includes public platforms like Twitter, Facebook, Instagram, Reddit etc.

1.1.2 Mental Health

An individual's mental health can be described as the state of mind that produces statisfaction, happiness, ability to adjust to stress e.t.c. Deterioarating mental health is often indicated by stress, depression, and rise in anger and violent urges[6]. Maintaining a good mental health has positive effects on the productivity and longevity of a person. On the contrary, bad mental health can lead to various diseases and cut the lifespan short.

1.1.3 Hate Speech

Any phrase expressing extreme hatred towards other people or self can be defined as hate speech. Hate speech can cause critical damage to the mental health of the person towards whom it is targeted. Hate speech that is targeted towards communities based on religion, gender can some times lead to violence between the concerned communities.[17]

1.1.4 Effects of Social Media on Mental Health

How users engage in discussion on public platforms like Facebook, Instagram, Twitter greatly affects users mental health[5]. With most young people using the internet for social networking thereby, exposed to vast content - suitable and non-suitable on regular basis. This has resulted in the rise of mental health concerns in young people over the past years [5], substantiating the link between social media and depression.

Need to reduce hate speech: A mechanism to warn us against posting hateful or hurtful comments will be beneficial for our collective well-being. Recent research on datasets collected from patient-doctor interactions on online health forums[2] have given us insight into mining of emotions relating to mental-health from long text sequences [12]. The same principles are being in the proposed solution and extend our dataset to social media.The proposed model will be trained to mine and classify emotions from public comments and posts on social media. This will help detect hate speech and warn against posting negative comments directly on the public forums.

1.2 Background

To understand the objectives of this project and the models used in the proposed solution used to fulfill the said objectives, following architectures play fundamental role.

Transformer — The Transformer model is well suited for NLP tasks as it handles dependencies between words/phrases that are not immediately next to each other but are separated by a lot of words in between. This is known as long-term dependency. The sequence- to -sequence learning paradigm of the transformer model helps it to learn context better than other previous models.

The transformer architecture is basically made of a stack of encoder layers and a stack of decoder layers. The first Encoder receives the input word embeddings. In each encoder layers transformation is carried out and passed to the next. At last the transformed embeddings (called context vectors) are passed to all the decoder layers parallely.

Finally the Decoder layer of the transformer outputs the target sequence or predicts the probabilities of the traget sequence. Self attention[28] (also knows as Query Key View attention) inside both the encoder and decoder stacks is the most important feature of this model and helps to form embeddings that give more attention to words that have more importance in the given context.

BERT[4] is a transformer-based machine learning technique for natural language processing (NLP) pre-training developed by Google.

RNN - RNN, stands for Recurrent Neural Network, belongs to the category of Artificial Neural Network (ANN) which is widely used in application areas / fields of Natural Language Processing (NLP) including Sentiment Analysis and Speech Recognition. An RNN model uses sequential data or time series data. It is designed to recognize the sequential characteristics of data and thereafter using the patterns to predict the coming scenario.

The class of networks having an infinite impulse response (IIR) is precisely referred using the term RNN - Recurrent Neural Network. An IIR can be defined as a directed cyclic graph which is not possible to be unrolled and they can possess additional stored states and that storage are subject to be under direct control by the neural network. These types of controlled states are often referred to as by the term

- gated state or gated memory and are part of network known as long short-term memory networks (LSTMs) and gated recurrent units, which alternatively is also called as Feedback Neural Network (FNN). Bidirectional long-short term memory (BiLSTM)[21] is a type of recurrent neural networks and is implemented by putting two independent RNNs together.

The models that used in this project are BERT and BiLSTM which are defined in later sections.

BERT– BERT (Bidirectional Encoder Representations from Transformers) is an NLP model released by researchers at Google AI Language[4].BERT's key technical innovation is applying the bidirectional training of *Transformer*[29], a popular *attention model*[29], to language modelling. This is in contrast to previous efforts which looked at a text sequence either from left to right or combined left-to-right and right-to-left training. The architecture of this model is described in *Figure1.1*.

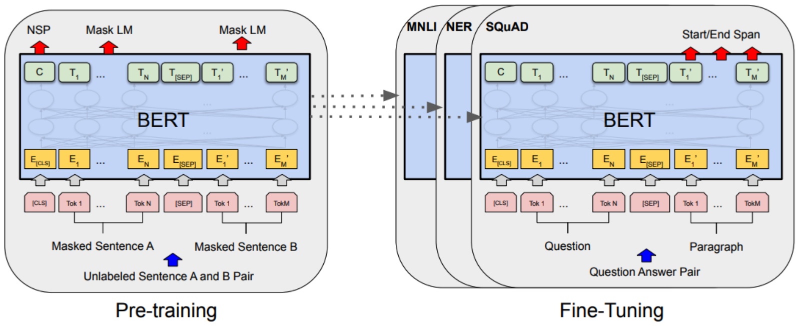

Figure 1.1: BERT architecture as implemented by Google AI Language researches[4]

Most conditional languages need deep bidirectional representation. This is achieved by training the model to predict *masked words*.Some percentage of the input tokens are merged at random, and then those masked tokens are predicted. This procedure is referred as a "masked LM" (MLM)[20]

In order to extract proper contextual meaning the model has to be trained to predict next sentences referred as Next Sentence Prediction(NSP)[26].Therefore pre-training of BERT is achieved through MLM and NSP. The next vital phase is the *fine-tuning* with supervised data (task specific inputs and outputs). Instead of independently encoding text pairs before applying bidirectional cross attention, BERT uses the self-attention[25] mechanism to unify these two stages, as encoding a concatenated text pair with self-attention effectively includes bidirectional cross attention between two sentences.

BiLSTM–A Bidirectional LSTM or BiLSTM[7] [22], is a sequence processing model that consists of two LSTMs: one taking the input in a forward direction, and the other in a backwards direction. It is a type of recurrent neural networks.

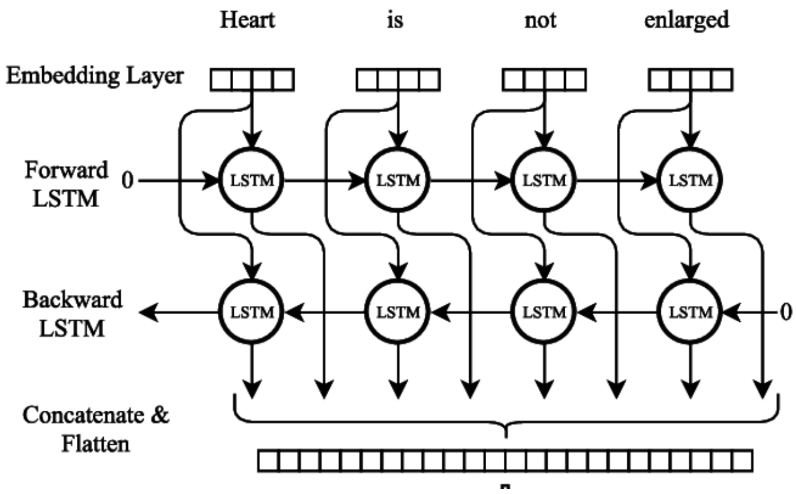

Figure 1.2: BiLSTM Architecture as used in [7]

As shown in *Figure 1.2*, BiLSTM structure allows the networks to have both backward and forward information about the sequence at every time step. This ensures in increase of the amount of information available to the network, improving the context available to the algorithm (e.g. knowing what words immediately follow and precede a word in a sentence).

PSO – Particle Swarm Optimization is a function optimization technique inspired by social behaviour of bird flock. It is proposed by Kennedy and Ebehart in 1995[11]. This algorithm helps in convergence of a function close to it's global minima.

This algorithm is based on the observation about how one bird allows all birds in the swarm to expand their observable vicinity. Each particle has an associated position, velocity, fitness value and keeps track of it's own best fitness value and best fitness position. Global best fitness value and global best fitness position for the swarm are also maitained.

The fitness of any function can be converged to a very close value near it's global minimum, thereby, we can use PSO[11] to optimize hyperparameters and fine-tune[27] different existing models.

PSO assumes a random population of m particles with each particle as potential solution to the problem that needs to be solved in the given search space. In a *d-dimensional* search space, suppose there are *m* particles, the velocity and position of *i*-th particle at the time t are expressed as

$$v_i(t) = [v_{i1}(t), v_{i2}(t), ..., v_{id}(t)]^T \tag{1.1}$$

$$x_i(t) = [x_{i1}(t), x_{i2}(t), ..., x_{id}(t)]^T \tag{1.2}$$

In each iteration, the positions and velocities of each particles are updated using two values:

- p_b : Best value of particle.
- g_b : Best value of population overall previous

The above two values at iteration *t* are:

$$p_{bi}(t) = [p_{i1}(t), p_{i2}(t), ..., p_{id}(t)]^T \tag{1.3}$$

$$g_b(t) = [g_1(t), g_i(t), ..., g_d(t)]^T \tag{1.4}$$

These position and velocity values are updated at iteration *t+1* as follows:

$$v_i(t+1) = wv_i(t) + c_1 r_1 (p_{bi}(t) - x_i(t)) + c_2 r_2 (g_b(t) - x_i(t)) \tag{1.5}$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{1.6}$$

- w : inertia weight coefficient, for trading off global against local search ability.
- c_1, c_2 : algorithm learning factors.
- if $c_1 = 0$, then easy to fall into local optimization and cannot jump out.
- if $c_2 = 0$, it leads to slow PSO convergence speed.
- r_1, r_2 : random variables uniformly distributed in [0, 1].

Only the optimal particle can transmit information to other particles in each iteration. The algorithm terminates when maximum number of iterations reached or a very good fitness value is obtained.

1.3 Objectives

- To identify and detect the presence of hate speech/quotation in the comments that are made on public platforms.
- Flagging comments for sad/negative emotions like Depression, Suicide etc. which may be resolved by suggesting proper psychiatrist and professional practitioners by means of popups and triggers.
- Hate speech comments are to be flagged by training our model from the public comments on social platforms and review platforms and let our model understand the natural emotions inside the texts.
- To explore the previously reviewed paradigms and aim to enhance the accuracy of the previous models so as to come up with better approaches to the same problem.
- Adding a transformer (Bi-LSTM) layer to BERT to increase accuracy of classification.
- To tune BERT by optimizing the BERT hyper-parameters using hyperparameter optimization technique (PSO) to get better accuracy over proposed model.

Chapter 2

Related Work

2.1 Effect of social media on mental health

Human beings are social animals and require constant connection with others to survive. Social media allows this connection with a vast variety of people all over the world. But, this connection provided by social media also comes with it's own share of mental health problems. So, a systematic review of all the available literature work to assess positive and negative effect social media can have on mental health,[5] concludes that there is no solid evidence of social media effecting mental health of people.

Users on social media can learn from the painful experiences of other people and find courage and self-esteem from these experiences. Social media can also be used as a platform to find self-identity and self-expression. All these effects can be considered positive. On the contrary, there are also negative effects like fear of missing out, anxiety, depression that can easily detoriate the mental health of the user.Thus, as stated in [23], social media should be used with utmost awareness.

2.2 Existing paradigms for Sentiment Analysis

With the growing influence of social media and various online review platforms, comments have become a very important way for people to express their opinions. By analyzing the sentiments of these comments, we can learn a lot of crucial information. [7] does exactly this by using the transformer model BiLSTM. This work[7] has achieved an F1-score of 92.18% indicating correct detection of sentimentns from the comments.

Text has become an important expression of opinion in this age. The sentiment analysis of massive inflow these text messages have huge monetary value in various industries. Doing sentiment analysis of text by reframing sentiment classification as a comparing problem in addition to using BiLSTM and MHA for feature extraction, [15] has achieved an accuracy of 89.5% on IMDB dataset.

Incase of a long-text sequence data, there exists ambiguity in emotions of the sequence. To improve effciency of detecting negative emotions in such cases, deep learning with transformers can be used[12]. By using MHA and applying BiLSTM and Convolutional Neural Network(CNN) over it, [12] has achieved an accuracy of 96.2% using the dataset of medical heathcare platform with patient-doctor interaction.

With the increasingly fast-paced world, people are having less time to connect face-to-face with others. This is increasing loneliness and negatively affecting their mental health and in extreme cases, this can lead to suicides. Detecting suicidal emotions using a lexicon based BiLSTM with MHA and CNN on top, [13] has achieved an accuracy of 97.8% on reddit dataset.

In twitter, like many other social media platfoms, vast amounts of information is being propogated through texts. It is challenging to extract sentiment from the text with dynamically changing internet slangs. Using BERT and finetuning it for sentiment analysis by adding a single layer on top, [24] achieved an accuracy of 87.5%

on twitter data.

Text data is exploding on various social media platforms with many comments being from investors and consumers. Sentiments of these investors and consumers have a profound effect on energy market. Analyzing these sentiments can reveal important information about the trends in energy market. By using BERT and fine-tuning it with a BiLSTM layer, [21] has produced an accuracy of 86.2% on the dataset consisting investors' and consumers' statements on chinese internet.

2.3 Finetuning BERT by combining with other models

Combining Bi-LSTM with BERT using statistical data to classify the sentiment/stance of consumers and investors, [21] gives better recall of 0.8620 compared to the recall of BERT model (0.855) and LSTM model (0.775).
A BERT+BiLSTM+textCNN efficiently extracts contextual information and perform multi-class classification of emotions [9]

[30] is based on the BERT model with certain key differences like:

(1) training the model longer, with bigger batches,over more data

(2) removing the next sentence prediction objective

(3) training on longer sequences

(4) dynamically changing the masking pattern applied to the training data This model achieves an accuracy of 83.2% compared to BERT accuracy of 72%.

Introducing a light weight version of BERT[31] where the training process is less computationally expensive and exhausts less memory.

2.4 Optimizing BERT through hyperparamter tuning

Fine-tuning a transformer model requires the most optimal set of hyperparameters like learning rate, batch-size, number of epochs, wieght decay etc. Hyperparameter Optimization techniques are often used along with language learning models for getting more efficiency. Using the Genetic Algorithm which is an evolutionary algorithm to optimize hyperparameters of BERT [19]yields better weighted average F1 scores.

This indicates that using a more modern algorithm for finding global optima for non-linear functions can be used. One such algorithm is the Particle Swarm Optimization[11] introduced in 1995. PSO along with Bayesian Optimization technique can be used for hyper-parameter estimation. This paper develops a new approach PSO-BO[14] which can be used to optimize the acquisition functions.

2.5 Hate Speech detection from videos

With the ample use of video sharing sites,and rapid spread of media containing hate speech through social media there is a need to find a way to detect hate speech in videos. All the current employed methods to flag a video are based on violence content in the video i.e in most cases only the visual evidence of violence is detected and any hate speech is going undetected. This [3] extracts audio and converts it to text through Speech to Text methods and then feeding it into machine learning models.

2.6 Using Multilingual dataset for training model

Comments or discussion on informal social media platforms are often multilingual.Hence a model trained solely on english might not be able to capture the full context of comment. Therefore, multi-lingual sentiment analysis is imperative. [16] reviews the difficulties of this task such as ambiguity, translation error and suggests recommendations. Google Research has released a model of bert for multi-lingual classification trained on 104 languages.

Chapter 3

Problem Statement

S_i	Source of data-set
E_i	Set of expressions present in text
T_i	Set of text data from source
x	Word vector $\{x_1, x_2, x_3 ...\}$
t	Time step
i	Input gate
Z	Activation vector for input
f	Forgetting gate
C	Candidate values
o	Output gate
h	Hidden states $\{h_1, h_2, h_3 ...\}$

The objective of the project as described in *Section 1.3* can be broadly divided into two major problems.

3.1 Hate Speech and Negative Emotion Detection Problem

As explained in *Section 1.1.3*, hate speech or negative emotions can affect mental health of self or other person which can cause change in behavior or generation of bad or negative thoughts. This can also lead to mental issues and taking life threatening or serious actions as described in *Sections 1.1.2 and 1.1.4*. Social media is one of the major cause for spread of hate speech leading to development of negative thoughts and emotions.

Statement :

Let $S = \{S_i\}$ be the set of source of data and data set to retrieve text T. The problem is to detect the emotion E expressed by the a text T_i.
Let $E = \{E = E_i\}$ be the set of all emotions that can be present in a textual data.

Let T_i represents a single text data taken from source $S_i \subset S$ that may contain emotion $E_i \subset E$. The task is to find E and categorize it in negative / hate / neutral / fear emotion etc.

The most effortless way to solve is feeding the certain set of texts called training data T_{train} into a deep learning model and test is against the other set of textual data called training data T_{test}.

The black box function M can be defined as the model learning over T_{train} and it returns the predictor P through which T_{test} is passed to check accuracy of the model.

$$P = M(T_{train}) \tag{3.1}$$

$$E = P(T_{test}) \tag{3.2}$$

Furthermore the model M can be visualised as as combination of models having multiple layers and several functions. T_{train} is passed to pre-trained BERT model to get word vector $x : \{x_1, x_2, x_3 ...\}$ from the input text. Formula given in 3.1 can be seen as series of functions:

$$
\begin{aligned}
&x = BERT\left[T_{train}\right] \\
&i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}C_{t-1} + b_i\right) \\
&Zt = \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \\
&f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right) \\
&c_t = f_t c_{t-1} + i_t z_t \\
&o_t = \tanh\left(W_{xo}X_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right) \\
&h_t = o_t \tanh\left(c_t\right)
\end{aligned}
\tag{3.3}
$$

The above relation indicates the core gates of LSTM - forgetting (f), input (i), output (o) gates.

For a comment or review posted on public platforms like social media or review website, this project deals with detection of emotions and presence of hate speech in these. Any textual content that may result in negative emotions like hate or anger for the reader should be detected and flagged.

This is accompanied by using the pre-trained model as mentioned in 3.1 to build a prototype to flag the targeted comments and reviews.

3.2 Hyper-parameter Estimation and Model Optimization problem

Text obtained from comments and reviews are fed into the model for detection of hate speech and negative emotions. This result will be achieved by combining pre-trained BERT model with BiLSTM model as discussed in [13] and [21].

The goal here is to estimate the hyper-parameter by fine tuning the model to optimize the result obtained from BERT + BiLSTM model. Hyper-parameters are those parameters whose values control the learning process and also responsible for determining the model parameters values that a learning algorithm evaluates to be learning. The optimization algorithm used here for tuning and estimation is Particle Swarm Optimization (PSO).

The hyper-parameters as mentioned above are defined in *Table 5.1*. The goal of this project is to estimate the combination which can optimize the model and maximize the accuracy obtained. Consider the *Equations 1.5 and 1.6* which determines and update the features to get local and global best results and hence making it an optimization problem.

Chapter 4

Design and Implementation

This chapter discusses the implementation of the model, the metrics for evaluating performance and shows the working prototype for emotion detection whose Back-end Server is connected to the proposed model.

BERT is a popular transformer model that can be used for hate speech detection. After the above literature review, it is known that finetuning BERT by adding a layer of other transformer models like BiLSTM and optimizing hyper-parameters of BERT can improve the accuracy with which hate speech is detected. The same possibility is being investigated in this report.

4.1 Existing Work

As discussed in *Section 2.2* extensive research has been performed in the field of sentiment analysis. Several methods have been adapted to get better and optimized results. Solutions proposed have used models like

- Bi-LSTM[7]
- BiLSTM + MHA[28]
- Bi-LSTM + MHA BCNN[12]
- BERT + Bi-LSTM[21]
- BERT + BiLSTM + textCNN[9].

It is observed that constructing combined architectures can give better results than stand-alone models. *2.4* discusses previous work related hyper-parameter tuning of BERT for more efficiency. Compared to naive Algorithms like Grid Search or Random Search, informed meta-heuristic search algorithm that take previous iterations into account perform better. The Genetic Algorithm has been previously used to optimize BERT [19]. This indicates scope for exploring other modern algorithms for hyper-parameter tuning.

4.2 Proposed Solution

Several models and techniques were explored as discussed above. Upon evaluation of various models we have come to the conclusion that PSO has not been applied to this use case till date. PSO can guarantee the convergence of the algorithm on choosing proper parameters within its own stability region. We have taken the BERT model with and without BiLSTM model to calculate the score and optimally deciding the hyper parameters for getting better score/accuracy.

In order to demonstrate the performance of the proposed method, we have analyzed mainly two data sets and compared the result with different kinds of combination in accordance with and without hyper parameter tuning.

4.2.1 Modified Method

BERT+BiLSTM outperforms the baseline BERT model [10] as shown in *Table 5.4*. This can be further optimized by tuning the hyper-parameters as described above in *Table 5.1*. The flow diagram shown in *Figure 4.2* shows the flow of data and how it is evaluated.

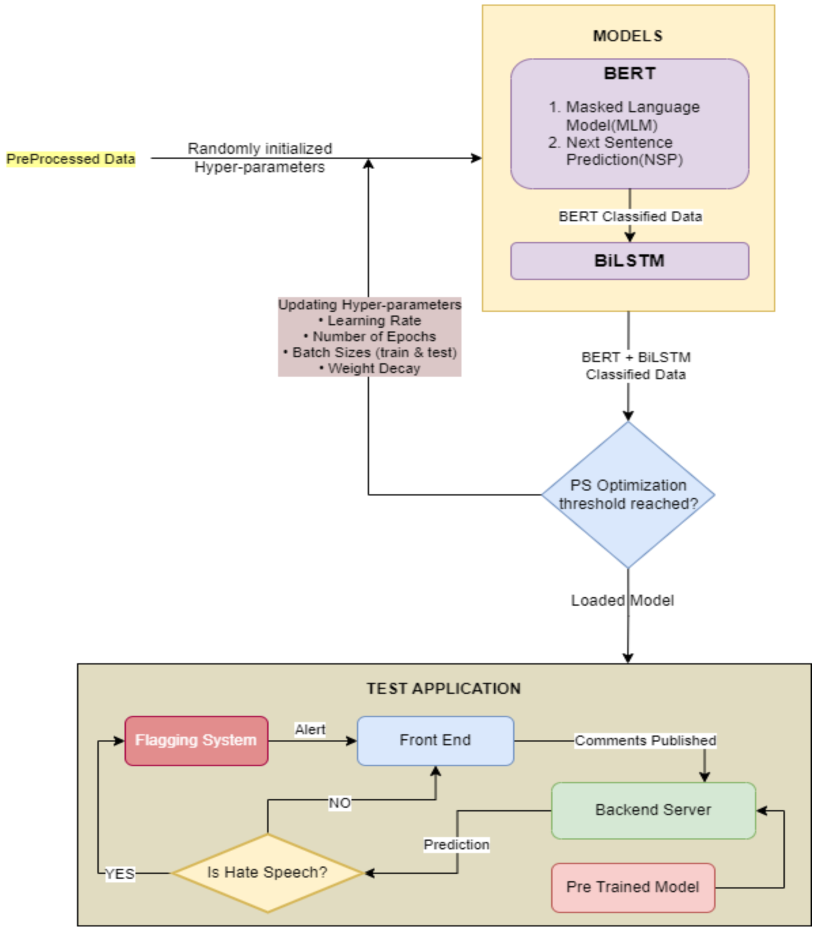

Figure 4.1: Architecture for the proposed model, BERT And BiLSTM Model hyper-tuned with PSO, along with High Level Design for flagging system

Architecture

Consider the Architecture of the project *Figure4.1*, the processed data is passed to our customized model, where the BERT acts as base model from where BERT clas-

sified data is obtained which is further fed into BiLSTM learning model for training and validations. This is the model architecture of the project.
The hyper-parameters are initialized for model combination as stated above. These hyper-parameters are tuned by Optimization Algorithm namely PSO[14] as discussed in previous section.

Hyper-tuning of Parameters

1. Initiate Hyper-Parameters: Unlike normal parameters (values that are learnt during training), hyper-parameters like learning rate, number of epochs, weight decay, batch size are the values that control the training process itself. We define a search space for this hyper-parameters as illustrated in *Table5.1*. For first iteration a set of hyper-parameters are randomly selected from this search space.

2. Learning Process through BERT+BiLSTM model: For each evaluation, we then select hyper-parameters and pass to the BERT+Bi-LSTM model. The model trains on the embedded data set fed to it and calculates train loss and train accuracy. The learning process continues for specified number of epochs.

3. Calculate accuracy: After the specified number of epochs are executed in the above step the learned model is used on the test dataset. The test accuracy are calculated on the pre-classfied test dataset. The fitness of the learned model is calculated using metrics mentioned *4.2.2*.

4. Update Hyper-Parameters: This step uses the Particle Swarm Optimization technique explained in *section1.2*. For each evaluation, PSO initializes a population of particles, where each particle represents a set of hyper-parameters. Each particle has associated velocity, position and inertia and also has knowledge of other particles in its neighbourhood. The attributes of the particles are updated according to *equation1.6* and *equation1.5*. This process ensures that optimal particle transfers its values to surrounding particles. Ultimately only the best particle remains. The attributes of the particle are the chosen hyper-parameters for next evaluation.

5. We perform the above steps until threshold is reached. Here we have evaluated the model 25 times each with hyper-parameters set by PSO method as mentioned in *Section 1.2*.

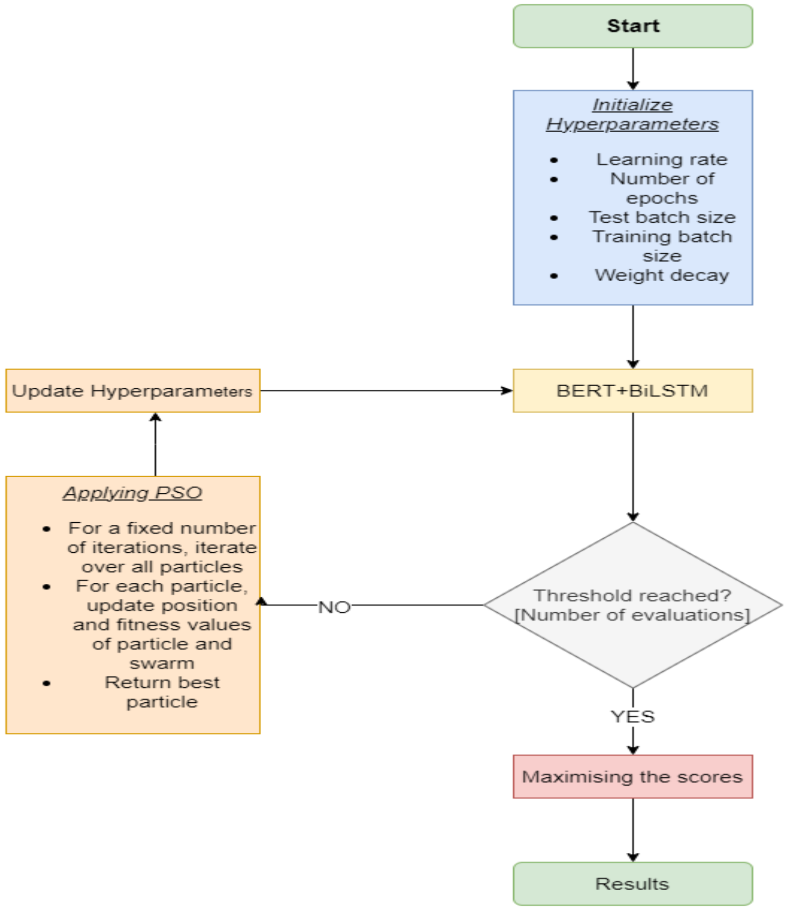

Figure 4.2: BERT And BiLSTM Model hyper-tuned with PSO

We have done a comparative study by doing experimentation with models like BERT, BERT+BiLSTM and also PSO Hyper-parameter tuning with BERT and BERT+BiLSTM.

All the scores / accuracy obtained are shown in *Table 5.4*. The results are plotted as graph and shown later in section.

4.2.2 Metrics for Evaluation

F1-score metric is used in deciding the best model to be used among the above combinations.

F1-score is an evaluation metric to assess predictive skill of a model. This metric combines two competing metrics-precision and recall scores of a model.

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4.1}$$

Precision is an evaluation metric that shows the proportion of positive identifications that were actually correct.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{4.2}$$

Recall is an evaluation metric that shows the proportion of actual positives that were identified correctly.

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{4.3}$$

More Precision involves a harsher critic doubting even actual positive samples and more recall involves a lax critic that can't classify border-case negative samples correctly.For this reason, we use F1-score that is a harmonic mean of both precision and recall as a metric for our model.

4.3 Working Prototype

The pre-trained model is used by Backend Server running on *Python Django* to process the comments obtained from the application UI and predictor works on it to decide the presence of hate speech in it, thereby alerts are made accordingly as shown in *Figure 4.3*. The comments are also flagged by different colours as shown in *Figure 4.4* to mark various types of emotions detected in the comment.

Using pre-trained model, the response time from server is observed as fast as 0.2 seconds (look *Table 4.1*), average being around 0.4 seconds shows clearly the improvement of building a learning model over existing BERT model.

Table 4.1: Server Response Time

Response Type	Response Time (in milliseconds)
Earliest	200
Average	400-500
Latest	2100

Figure 4.1 shows the architecture of our application how the pre-trained model is loaded and used for flagging the comments.

Figures 4.3 and 4.4 demonstrate our working prototype for the solution we proposed in this project - flagging and detecting hate / negative emotions in texts on social media.

After getting the optimized hyper parameters, the model is loaded into our back-end server. We use pre-loaded model for this use case to reduce latency in real time interactions. On commenting any text, before posting on to the database, the back-end server runs a test over the obtained text and the prediction is stored in database and same is returned back to application for raising flag if required as shown in Figure 4.4.

```
[01/Mar/2023 05:21:23] "GET /api/post/1? HTTP/1.1" 200 7068
1/1 [==============================] - 0s 443ms/step
{'message': 'I am playing', 'emotionDetected': 'neutral', 'raiseAlert': False}
[01/Mar/2023 05:21:38] "GET /api/predict?text=I%20am%20playing&postId=1 HTTP/1.1" 2
[01/Mar/2023 05:21:38,096] - Broken pipe from ('127.0.0.1', 53960)
[01/Mar/2023 05:21:38] "GET /api/post/1? HTTP/1.1" 200 7222
1/1 [==============================] - 0s 448ms/step
{'message': 'I am killing animals', 'emotionDetected': 'fear', 'raiseAlert': True}
```

Figure 4.3: Server Responses with raiseAlert:False for neutral and positive comments and raiseAlert:True for negative comments

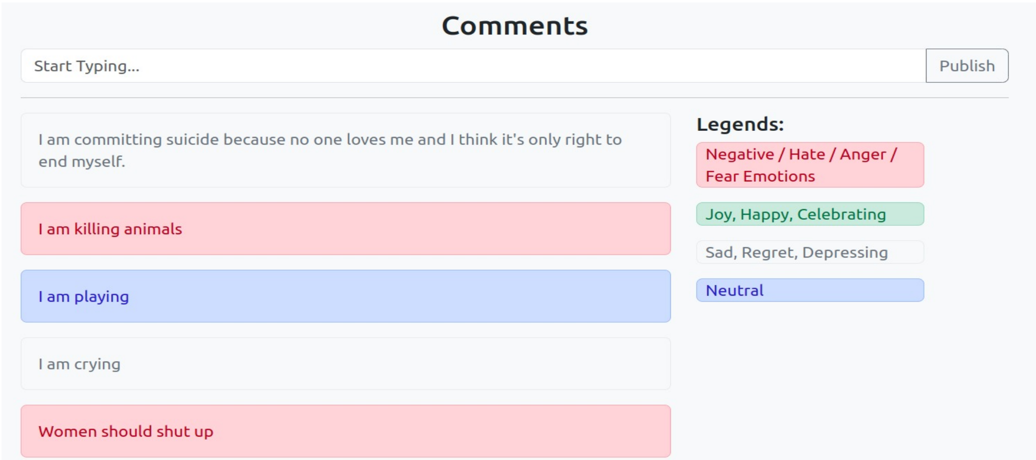

Figure 4.4: UI of our Application showing green for Positive, red for Negative and Blue for Neutral comments input by user

Chapter 5

Results and Discussion

5.1 Data Set

This section talks about the data set used in the proposed model and how the data points are set up against it.

5.1.1 Description

Two major data sets are used in this experiment. Tweets dataset and IMDB reviews datasets from Kaggle. Both the data sets are labelled. Tweets data set contains 11,328 examples (test & train) [8]. The IMDB dataset from kaggle called the Large Movie Review Dataset[18] is a collection highly polar movie reviews.In both the

Table 5.1: Type and Range of hyper-parameters for BERT + BiLSTM

Name	Type	Range
Learning rate	Real	$(1 \times 10^{-5}, 1 \times 10^{-3})$
Number of epochs	Integer	(30, 50)
Train batch size	Integer	(1, 5)
Test batch size	Integer	(10, 20)
Weight decay factor	Real	(0.005, 0.015)

cases, each example has 2 dimensions for classification procedure.

5.1.2 Setup

We first tried to identify the major hyper parameters for the proposed BERT + BiLSTM model. We estimated 5 such hyper parameters - learning rate, number of epochs, train batch size, test batch size and decay factor. The experiment was ran for 25 times and accuracy obtained from BERT and BERT + BiLSTM models was considered as score for finding optimal hyper parameters. *Table 5.1* shows the estimated hyper-parameters, their types and their allowed ranges. This is fine tuned to get optimized result.

The hyper-parameters are initialized as per the ranges described in *Table 5.1* and model is fed with pre-processed data along side these hyper-parameters. After the model accuracy is obtained, the PSO is implemented over these hyper-parameters to get new values. All the values are updated and the process goes on until a threshold is reached. Here we have taken 25 evaluation rounds as threshold limit. Accuracy obtained is maximized to get optimal configuration or hyper-parameters.

5.2 Experimental Results

- Performed the hyper-parameter estimation / tuning for BERT and BERT + BiLSTM models using particle swarm optimization. This hyper-parameter tuning method is the novelty of the project presented. This is done as part of experimental research rather than optimal research pattern.

- Hyper-tuning was performed as explained above in *Experiments* Section. The hyper-parameters obtained after tuning BERT+BiLSTM at each evaluation are shown in *Table 5.2*.

- For the sake of research and comparative study with the above experiment, we have also included the accuracy obtained from running BERT and BERT+BiLSTM models without hyper parameter tuning. All the accuracy / score obtained for BERT (*Figure 5.1*) and BERT+BiLSTM are shown in *Table 5.4*

```
107/107 [==============================] - 98s 865ms/step
              precision    recall  f1-score   support

         joy       0.85      0.84      0.85       707
     sadness       0.81      0.83      0.82       676
        fear       0.89      0.84      0.86       679
       anger       0.79      0.80      0.80       693
     neutral       0.80      0.82      0.81       638

    accuracy                           0.83      3393
   macro avg       0.83      0.83      0.83      3393
weighted avg       0.83      0.83      0.83      3393

array([[597,  15,  16,  16,  63],
       [ 17, 559,  16,  58,  26],
       [ 23,  31, 571,  43,  11],
       [ 19,  60,  28, 555,  31],
       [ 43,  29,  11,  29, 526]])
```

Figure 5.1: BERT Model prediction and accuracy

- *Figure 5.2* shows the plot of accuracy (Y-axis) against the learning rate(X-axis) of each PSO evaluation. The learning rate is expressed in the order of 10^{-5}. There is an observed trend of decreasing accuracy as the learning rate appoaches 10^{-4}. The plot shows that best accuracy has been achieved with learning rate of 0.296×10^{-5} in the sixth evaluation as illustrated in *Table 5.2*

- *Figure 5.3* consists of two plots. One of them represents the accuracy of classical BERT model without PSO the other represents the accuracy of BERT combined with Bi-LSTM without PSO plotted against number of epochs on the X-axis. It is observed that The BERT+Bi-LSTM model achieves consistently better accuracy as illustrated in *Table 5.4*

- The highest accuracy obtained after fine tuning the hyper-parameters of BERT+BiLSTM model by PSO is 93.1% which is better than maximum accuracy obtained on running BERT+BiLSTM model with manually initialized parameters. The final hyper-parameters obtained as optimal configuration as

Table 5.2: Hyper-parameters obtained on tuning BERT+BiLSTM by PSO

accuracy	**epochs**	**weight decay**	**learning rate**	**train batch size**
86.1	30	0.0118	2.98×10^{-5}	4
50.4	45	0.0068	7.48×10^{-5}	2
53.8	22	0.0093	9.73×10^{-5}	1
52.9	39	0.0107	9.71×10^{-5}	1
52.1	32	0.0132	7.45×10^{-5}	2
93.1	**47**	**0.0082**	$\mathbf{2.96}\times10^{-5}$	**4**
86.3	22	0.0089	2.39×10^{-5}	2
50.4	37	0.0139	6.89×10^{-5}	4
57.2	30	0.0114	9.14×10^{-5}	3
52.9	45	0.0064	4.64×10^{-5}	1
84.6	34	0.0101	3.52×10^{-5}	4
70.0	22	0.0099	6.5×10^{-5}	1
91.4	37	0.0149	2.01×10^{-5}	3
88.0	23	0.0102	1.72×10^{-5}	4
57.2	38	0.0052	6.22×10^{-5}	2
57.2	30	0.0077	8.4×10^{-5}	1
76.0	45	0.0127	3.97×10^{-5}	3
53.8	44	0.0142	9.72×10^{-4}	4
52.9	40	0.0129	3.54×10^{-4}	3
54.7	32	0.0054	6.09×10^{-4}	2
53.8	47	0.00104	1.06×10^{-4}	4
52.1	34	0.0089	5.10×10^{-5}	2
62.3	49	0.0139	9.60×10^{-5}	4
57.2	27	0.0114	7.35×10^{-5}	3
57.2	42	0.0064	2.85×10^{-5}	1

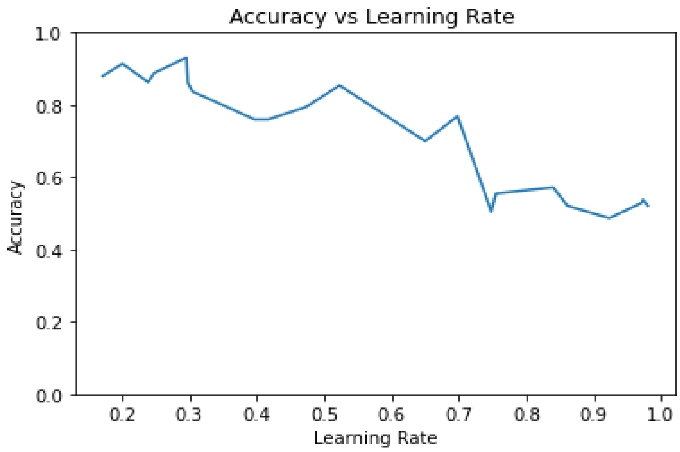

Figure 5.2: Accuracy plot as obtained by Hypertuning BERT+BiLSTM model with PSO

Table 5.3: Final hyper-parameter values obtained at maximum accuracy

Hyper-parameters	Value
Learning rate	2.96×10^{-5}
Number of epochs	47
Train batch size	4
Test batch size	14
Weight decay factor	0.0082

shown in *Table 5.3*.

- *Figure 5.4* shows the 3-D plot of *Accuracy* against *Number of Epochs and Learning rate*.

  As these are hyper-parameters, they are scattered in the 3-dimensional space. X-axis represents *number of epochs*, Y-axis represents *learning rate* and on Z-axis, *accuracy* is represented. This can be imagined as function $z = f(x,y)$. *Figure 5.2* shows how the accuracy is varied with one of the hyper-parameters - *Learning rate*. The maximum accuracy obtained on fine tuning the hyper-parameters using PSO is 93.1%, which is an improvement from the base model (BERT + Bi-LSTM) without hyper-tuning.

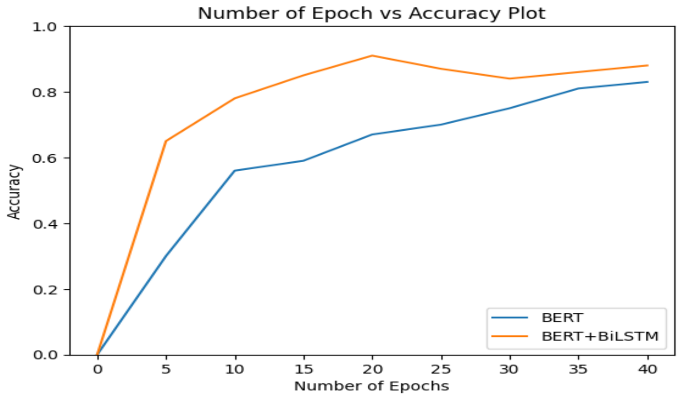

Figure 5.3: Accuracy plot as obtained by BERT and BERT+BiLSTM models without PSO hyper-parameter tuning

Table 5.4: Accuracy Comparison

Models	Accuracy/Scores
BERT	0.830
BERT with PSO tuning	0.882
BERT + BiLSTM	0.901
BERT + BiLSTM with PSO tuning	0.931

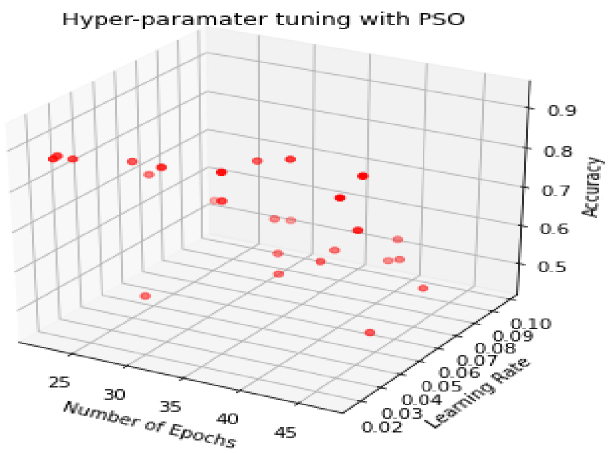

Figure 5.4: BERT+BiLSTM+PSO Model accuracy

Chapter 6

Conclusion and Future Work

6.1 Conclusion

It is observed that finetuning hyper-parameters of BERT using PSO inevitably increased the accuracy. Adding a layer of BiLSTM over BERT resulted in increase in accuracy. Based on the above two observations, hyper-parameters of BERT was fine-tuned and a Bi-LSTM layer was added on top of it resulting in improved accuracy.

6.2 Future Work

The current proposed solution has further scope of optimization by using a larger search-space for hyperparameters. Larger search space avoids the possibility of local minima.

The model can give improved results with larger dataset collected from web-scraping social media. Comments and posts from social media can provide greater fine-tuning for sentiment analysis.

The model can be extended by using multi-lingual text dataset. Before passing the data from non-english datasets to our model,it will be first converted to English from the given language as BERT works best on English Language.

As of now, videos posted on social media are only flagged based on graphic content and if reported by users. Hate speech detection and flagging from videos is the need of the hour. Videos extracted from social media and passed through speech to text conversion model can be fed to this model.

Bibliography

[1] Thomas Aichner, Matthias Grünfelder, Oswin Maurer, and Deni Jegeni. Twenty-five years of social media: a review of social media applications and definitions from 1994 to 2019. *Cyberpsychology, behavior, and social networking*, 24(4):215–222, 2021. 1.1.1

[2] Ya-Mei Bai, Chao-Cheng Lin, Jen-Yeu Chen, and Win-Chien Liu. Virtual psychiatric clinics. volume 158, pages 1160–a. Am Psychiatric Assoc, 2001. 1.1.4

[3] Unnathi Bhandary Ching Seh Wu. Detection of hate speech in videos using machine learning. 2020. 2.5

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. volume 1, page 4171–4186, 2019. (document), 1.2, 1.1

[5] Lamis F. Abdalla Reem Chaudhry Ehsanullah Safeera Khan Fazida Karim,

Azeezat A. Oyewande. Social media use and its connection to mental health: A systematic review. 2020. 1.1.4, 2.1

[6] Silvana Galderisi, Andreas Heinz, Marianne Kastrup, Julian Beezhold, and Norman Sartorius. Toward a new definition of mental health. *World psychiatry*, 14(2):231, 2015. 1.1.2

[7] Xiaoyu Qiu Ziheng Yu Xu Wu Guixian Xu, Yueting Meng. Sentiment analysis of comment texts based on bilstm. 2019. (document), 1.2, 1.2, 2.2, 4.1

[8] SHERIF (2021) HUSSEIN. “twitter sentiments dataset”, mendeley data, v1. 5.1.1

[9] Xuchu Jiang, Chao Song, Yucheng Xu, Ying Li, and Yili Peng. Research on sentiment classification for netizens based on the bert-bilstm-textcnn model. *PeerJ Computer Science*, 8:e1005, 2022. 2.3, 4.1

[10] Aneesh Komanduri† Justin Zhan† Kate Pearce, Tiffany Zhan. A comparative study of transformer-based language models on extractive question answering. 2021. 4.2.1

[11] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, volume 4, pages 1942–1948 vol.4, 1995. 1.2, 2.4

[12] Dheeraj Kodati and Ramakrishnudu Tene. Negative emotions detection on online mental-health related patients texts using the deep learning with mha-bcnn model. In *Expert Systems with Applications*, 2021. 1.1.4, 2.2, 4.1

[13] Dheeraj Kodati and Ramakrishnudu Tene. Identifying suicidal emotions on social media through transformer-based deep learning. pages 1573–7497, 2022. 2.2, 3.2

[14] Yaru Li, Yulai Zhang, Gongxue Zhou, and Yifei Gong. Bayesian optimization with particle swarm. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6, 2021. 2.4, 4.2.1

[15] Yuan Lin, Jiaping Li, Liang Yang, Kan Xu, and Hongfei Lin. Sentiment analysis with comparison enhanced deep neural network. volume 8, pages 78378–78384, 2020. 2.2

[16] Siaw Ling Lo, Erik Cambria, Raymond Chiong, and David Cornforth. Multilingual sentiment analysis: from formal to informal and scarce resource languages. *Artificial Intelligence Review*, 48:499–527, 2017. 2.6

[17] Abdurahman Maarouf, Nicolas Pröllochs, and Stefan Feuerriegel. The virality of hate speech on social media. *arXiv preprint arXiv:2210.13770*, 2022. 1.1.3

[18] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. 5.1.1

[19] Kosisochukwu Judith Madukwe, Xiaoying Gao, and Bing Xue. A ga-based approach to fine-tuning bert for hate speech detection. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 2821–2828, 2020. 2.4, 4.1

[20] Debora Nozza, Federico Bianchi, and Dirk Hovy. What the [mask]? making sense of language-specific bert models, 2020. 1.2

[21] YANGKEN CHEN LIANG ZHANG RUIJIANG YANG SHIWEI CHEN 1 2 REN CAI, BIN QIN and 2 WEI WANG 1. Sentiment analysis about investors and consumers in energy market based on bert-bilstm. 2020. 1.2, 2.2, 2.3, 3.2, 4.1

[22] Maryem Rhanoui, Mounia Mikram, Siham Yousfi, and Soukaina Barzali. A cnn-bilstm model for document-level sentiment analysis. volume 1, pages 832–847, 2019. 1.2

[23] Hassan Ebrahimpour Sadagheyani and Farin Tatari. Investigating the role of social media on mental health. *Mental Health and Social Inclusion*, 25(1):41–51, 2021. 2.1

[24] Ayman Samir, Saleh Mesbah Elkaffas, and Magda M. Madbouly. Twitter sentiment analysis using bert. In *2021 31st International Conference on Computer Theory and Applications (ICCTA)*, pages 182–186, 2021. 2.2

[25] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang. Bi-directional block self-attention for fast and memory-efficient sequence modeling, 2018. 1.2

[26] Wei Shi and Vera Demberg. Next sentence prediction helps implicit discourse relation classification within and across domains. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5790–5796, Hong Kong, China, November 2019. Association for Computational Linguistics. 1.2

[27] Tiancheng Tang, Xinhuai Tang, and Tianyi Yuan. Fine-tuning bert for multi-label sentiment analysis in unbalanced code-switching text. volume 8, pages 193248–193256, 2020. 1.2

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 1.2, 4.1

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 1.2

[30] Naman Goyal Jingfei Du Mandar Joshi Danqi Chen Omer Levy Mike Lewis Luke Zettlemoyer† Veselin Stoyanov Yinhan Liu, Myle Ott. Roberta: A robustly optimized bert pretraining approach. 2019. 2.3

[31] Sebastian Goodman Kevin Gimpel Piyush Sharma Radu Soricut Zhenzhong Lan, Mingda Chen. Albert: A lite bert for self-supervised learning of language representations. 2019. 2.3

PSO Hyper-tuning Project Plagiarism Report

ORIGINALITY REPORT

13%	12%	9%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	arxiv.org Internet Source	2%
2	github.com Internet Source	1%
3	Submitted to Visvesvaraya Technological University, Belagavi Student Paper	1%
4	Submitted to University of Wollongong Student Paper	<1%
5	sunghee.kaist.ac.kr Internet Source	<1%
6	www.rebellionresearch.com Internet Source	<1%
7	thequickadvisor.com Internet Source	<1%
8	www.researchgate.net Internet Source	<1%
9	turcomat.org Internet Source	<1%

10	Submitted to Moon Valley High School Student Paper	<1%
11	www.educba.com Internet Source	<1%
12	ipfs.io Internet Source	<1%
13	S. Mostafa Mousavi, Weiqiang Zhu, Yixiao Sheng, Gregory C. Beroza. "CRED: A Deep Residual Network of Convolutional and Recurrent Units for Earthquake Signal Detection", Scientific Reports, 2019 Publication	<1%
14	Shuhong Shi, Takahiro Uchiya, Ichi Takumi. "Emotion Recognition of Textual Session Record Based on the Voice Dialogue System of Reminiscence Therapy", 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE), 2021 Publication	<1%
15	spiral.imperial.ac.uk Internet Source	<1%
16	stanley.pdifashion.com Internet Source	<1%
17	ar5iv.labs.arxiv.org Internet Source	<1%

repositori.uji.es

18	Internet Source	<1 %
19	"Advances in Engineering Design", Springer Science and Business Media LLC, 2021 Publication	<1 %
20	gradesfixer.com Internet Source	<1 %
21	www.wenjiangs.com Internet Source	<1 %
22	ir.lib.uwo.ca Internet Source	<1 %
23	"Natural Language Processing and Chinese Computing", Springer Science and Business Media LLC, 2019 Publication	<1 %
24	dokumen.pub Internet Source	<1 %
25	jeannekoerber.com Internet Source	<1 %
26	Submitted to Aston University Student Paper	<1 %
27	kipdf.com Internet Source	<1 %
28	ogma.newcastle.edu.au Internet Source	<1 %

29	publications.lib.chalmers.se Internet Source	<1 %
30	web.archive.org Internet Source	<1 %
31	"Microelectronics, Electromagnetics and Telecommunications", Springer Science and Business Media LLC, 2019 Publication	<1 %
32	amtaweb.org Internet Source	<1 %
33	irlab.science.uva.nl Internet Source	<1 %
34	techblog.realtor.com Internet Source	<1 %
35	www.mdpi.com Internet Source	<1 %
36	"Intelligent Information and Database Systems", Springer Science and Business Media LLC, 2021 Publication	<1 %
37	Kian Long Tan, Chin Poo Lee, Kian Ming Lim, Kalaiarasi Sonai Muthu Anbananthen. "Sentiment Analysis with Ensemble Hybrid Deep Learning Model", IEEE Access, 2022 Publication	<1 %

38	Viktar Atliha. "Improving image captioning methods using machine learning approaches", Vilnius Gediminas Technical University, 2023 Publication	<1 %
39	compstat-lmu.github.io Internet Source	<1 %
40	dblp.uni-trier.de Internet Source	<1 %
41	eprints.lancs.ac.uk Internet Source	<1 %
42	i30www.ira.uka.de Internet Source	<1 %
43	repository.cityu.edu Internet Source	<1 %
44	www.iitp.ac.in Internet Source	<1 %
45	www.ijert.org Internet Source	<1 %
46	www.springerprofessional.de Internet Source	<1 %
47	Yaru Li, Yulai Zhang, Gongxue Zhou, Yifei Gong. "Bayesian Optimization with Particle Swarm", 2021 International Joint Conference on Neural Networks (IJCNN), 2021	<1 %

Publication

48	"PRICAI 2021: Trends in Artificial Intelligence", Springer Science and Business Media LLC, 2021 Publication	<1 %

Exclude quotes	On	Exclude matches	< 4 words
Exclude bibliography	On		