

Multiclass Text Classification Using Token-Level Learning Rate Decay In Bert

Pooja Pandey^{*1}, Dr. Sharvari Govilkar²

¹Research Scholar, Computer Engineering Department, Pillai College of Engineering, New Panvel , Navi Mumbai – 410210, Pooja1711.pandey@gmail.com

²Research Guide and Professor, Computer Engineering Department, Pillai College of Engineering, New Panvel , Navi Mumbai – 410210, sgovilkar@mes.ac.in

Abstract: Text classification is a procedure of categorizing a text into classes or predefined labels, allowing effective analysis and organizing the large volume of textual data. In the growth of online textual data like news, digital platforms, or social media, an effective model for text summarization and classification. However, traditional models of text classification had difficulty capturing deep semantic dependencies and contextual embeddings among texts or words, which led to misclassification and minimized classification accuracy. To overcome these challenges, the research proposed Token Level Class Importance (TLCI) and Layer Wise Learning Rate Decay (LWLRD) named as TLLRD is integrated with Bidirectional Encoder Representations from Transformers (BERT) named as (TLLRD-BERT), which is utilized for multiclass violence text classification and enhanced classification performance. The model integrates LLRD to stabilize fine-tuning across transformer layers and TLCI to emphasize semantically significant words associated with aggressive or violent expressions. The proposed TLLRD-BERT attains an accuracy of 98.24%, 96.54% and 99.71% through YouTube comments dataset, the news headline dataset and the SMS spam detection dataset, respectively.

Keywords: Bidirectional Encoder Representations from Transformers, contextual embeddings, Layer-wise Learning Rate Decay, Text classification, Token-level Class Importance, social media.

1. Introduction

In recent decades, the world faced an information overload crisis due to the vast amount of data generated across news platforms, search engines, social media and messaging services [1]. These sources produce significant volumes of unstructured text data every second [2]. As a result, modern machines utilise Natural Language Processing (NLP) to analyse and process text information [3]. The NLP enables computers to handle natural language data, which provides a consistent way to maintain the complexity and diversity of human languages [4]. A key objective of NLP is automatic text summarisation, which involves either extracting essential content from the original text or generating new content that focuses on the main ideas [5]. By compressing and refining the text, users are offered a concise summary. Due to manual summarisation being time-consuming and monotonous, automating this process has become enhancing popular [6]. Automatic summarization is performed on either a single document or multiple documents that address the same topic [7].

In recent years, Financial News Summarization (FNS) has emerged as a promising research area [8]. Thus, the NLP aims to reduce the length of financial news articles while preserving important events, emotions, and opinions that influence market fluctuations [9]. Hence, the proposed model based on the Deep Learning (DL) method is used for classifying abstractive summarization of news articles in English text [10]. The text detection task primarily focuses on recognising and localizing text regions within natural scene images. Several traditional approaches such as edge-based, texture-based, stroke-based, and connected component methods, are widely used in the violence detection domain [11]. In recent years, various Deep Learning (DL) approaches, such as

Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and other neural architectures, have been employed to automatically learn robust feature representations from large datasets [12]. Leveraging these capabilities, numerous DL-based text detection frameworks have been developed to enhance detection performance. DL-driven methods mainly aim to improve accuracy by utilizing deeper and more complex network structures [13].

- A Transformer-based model that incorporates Layer-wise Learning Rate Decay and Token-level Class Importance Weighting for improved contextual learning and token sensitivity.
- Ensures stable fine-tuning by maintaining lower-layer linguistic representations while adapting upper layers to task-specific semantics.
- Introduces token-level attention weighting to emphasise words that contribute strongly to violent or nonviolent cues.

This research paper is arranged in the following way: Section 2 presents the literature review, and Section 3 details the proposed methodology. Section 4 represents the experimental results, and the conclusion of this research paper is given in Section 5.

2. Literature Review

Anand [14] implemented an ensemble architecture of Bi-directional Long Short-Term Memory (Bi-LSTM) with naïve Bayes with hybrid Support Vector Machine (SVM) was utilized to distinguish unwanted content. A preprocessing stage involved filtering, noise removal and stop word removal and segmentation was performed first. In the feature selection stage, a Fuzzy-based Convolutional Neural Network (FCNN) was applied to select the relevant features. An ensemble architecture of Bi-LSTM with SVM was used to extract selected features and classification.

Yue [15] developed a knowledge fusion Network (KnowleNet) for multimodal sarcasm detection that leverages ConceptNet knowledge. It associates prior knowledge to increase sarcasm detection, though performed at the sample and word level, using modal semantic similarity analysis. To detect a semantic consistency of various modal data that learned useful features for sarcasm detection to increase the spatial distribution of sarcastic and non-sarcastic.

Ghanem and Erbay [16] introduced Contextualized B-LSTM (CBLSTM) for deep contextualized word representation. The CBLSTM model was integrated with a language model to address the spam texts issue on social networks. In the preprocessing stage, such as eliminating extra spaces, punctuations, URLs, special characters and emojis, stop words and converting a text to lowercase were performed to eliminate unwanted content in the text.

Vinoth and Prabhavathy [17] developed an Intelligent Machine Learning based Sarcasm Detection and Classification (IMLB-SDC) approach to detect sarcasm in social media. An IMLB-SBC approach were consisted of various operations like preprocessing, feature extraction, feature selection, classification and parameter tuning. A Term Frequency-Inverse Document Frequency (TF-IDF) was performed in the feature extraction phase, while chi-square and information gain were applied in the feature selection stage. An IMLB-SDC model consisted of SVM and C was optimally tuned through a Particle Swarm Optimisation (PSO).

Gedela [18] presented voting-based ensemble model for sarcasm detection by a DL model, while majority voting was performed on the results of four models to distinguish between sarcastic and non-sarcastic texts. A bidirectional encoder representation from a transformer to construct contextual word embeddings and fed to a network, which was an ensemble of DL models like CNN, Bi-LSTM, and parallel and sequential integrations of both were between models.

Al-Kabbi [19] introduced a CNN+LSTM integrated model that was utilized for classification, with a dual result was spam or not spam attained through a dense layer with one neuron and a sigmoid activation function. A punctuation removal, stop word removal, tokenization, and bag of words were preprocessing stage goal at prepare a raw message for efficient text analysis. In the feature extraction stage of CNN, LSTM and TF-IDF models were utilized to generate an integrated set of features that captured the temporal, local and global features of SMS messages. An operator operated a data fusion model to connect a feature from a feature extraction stage in a feature fusion step. The features achieved from every model were integrated to generate a more comprehensive and efficient representation of SMS data.

Thaokar [20] developed a fusion approach of CNN LSTM and CNN bi-LSTM to recognise a significant feature in a text to improve the classification of model performance. An N-gram relies on sarcasm detection for news and social media text through DL models, while necessary stop words, negation words and punctuation symbols were considered as various types of text features.

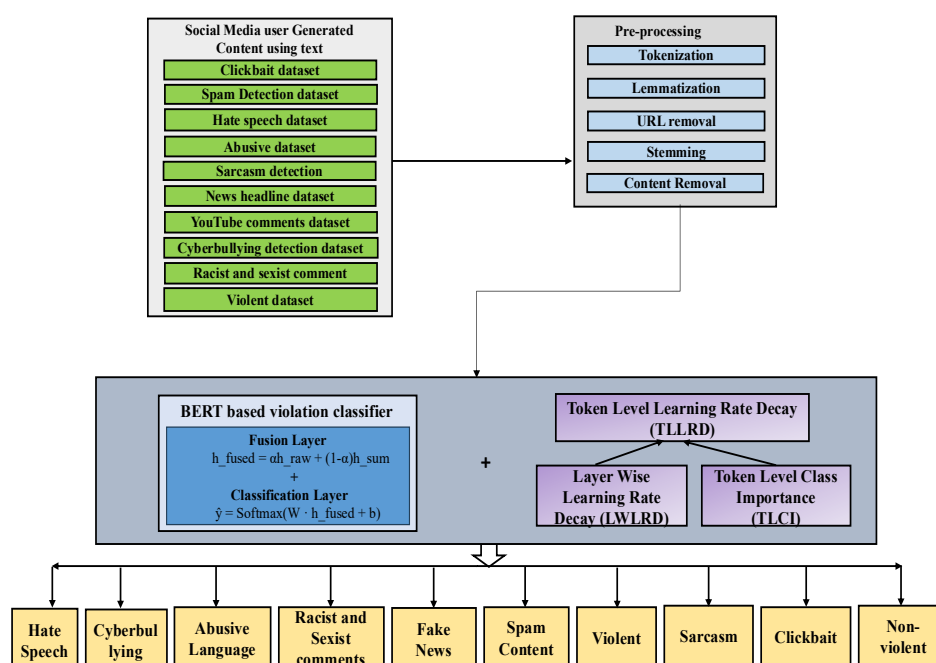


Figure. 1 Architecture of proposed TLLRD-BERT based multiclass text classification integrated with LWLRD and TLCI

One hot encoding feature vector was constructed and passed into the CNN architecture for input.

Arunthavachelvan [21] implemented a News Content and Comments (NCC) classification model, which utilized the psychological features extracted from the Psycho-Linguistic News Content and Comments (PL-NCC) dataset. By integrating the feed-forward layer into a DL model, which enhanced the effectiveness of the extracted features for detecting accurate fake news.

Daraghmi [22] developed an Integrated CNN-BiLSTM-GRU Model for Arabic Cyberbullying Detection from text to insight. In the preprocessing stage, text cleaning, data transformation, tokenization, stemming and vectorisation were performed to enhance the model performance. The CNN, Bi-LSTM, and GRU for text cyberbullying classification each of these models was unique characteristics and performance attributes. Aliyeva and Yaganoglu [23] introduced Multi-Layer Perception MLP was

performed for a feature fusion-based cyberbullying detection. A preprocessing step, like tokenization, stop word removal and text cleaning, was performed. In the feature extraction stage such as TF-IDF was performed to extract the necessary features.

3. Proposed Methodology

In this research, the proposed TLLRD-BERT is utilised to classify the text data into multiple classes in social media text. Social media text datasets include clickbait, spam detection, hate speech and abusive language, sarcasm, news headlines, YouTube comments and SMS spam detection. The above-mentioned classes are undergone in the preprocessing stage, like tokenization, stemming, token filtering and stop word removal, while a preprocessed text is fed into a BERT based classifier. The BERT classifier is performed by integration of LWLRD and TLCL for constant fine-tuning to emphasise critical linguistic cues. A model detects various violations such as clickbait, abusive language, sarcasm, violence, spam detection, hate speech, cyberbullying, fake news, racist and sexist comments, which are further validated through a BERT classifier estimator to ensure reliability and accuracy. Fig. 1 shows an architecture of the proposed TLLRD-BERT based multiclass text classification integrated with LWLRD and TLCL.

3.1 Dataset Description

The data are collected from different sources and included as nine classes such as. The sources are described below in detail:

Clickbait dataset [24]: The Clickbait Challenge dataset was employed, consisting of approximately 38,000 social media posts and article headlines from Twitter, annotated with clickbait intensity scores. This dataset aids in identifying exaggerated or curiosity-inducing headlines.

Hate speech and abusive language dataset [25]: The Hate Speech and Abusive Language dataset contains 24,783 Twitter posts categorised as hate speech, offensive language or neither.

Sarcasm detection [26]: A Sarcasm Detection dataset contains over 28,000 news headlines labelled as sarcastic or non-sarcastic. The headlines were collected from The Onion (sarcastic) and HuffPost (non-sarcastic). It captures subtle linguistic cues and irony, making it suitable for training models in sarcasm and sentiment detection.

News headline dataset [27]: The dataset consists of 28,619 headlines, of which 13,635 are labelled as sarcastic and 14,984 are non-sarcastic. The goal is to determine whether a headline is sarcastic without considering any URL. The dataset consists of no null values shows that it is complete.

YouTube comments dataset [28]: The YouTube comments dataset is split into two classes such as spam or ham. A dataset consisted of 1879 comments, which is categorised into 941 spam comments and 935 ham comments.

SMS spam detection dataset [29]: A spam detection dataset consists of 5574 SMS spam in which 4827 are denoted a ham and 747 a spam messages.

For the proposed experimental scenario of multiclass classification with 9 classes such as hate speech, cyberbullying, abusive language, racist and sex comments, fake news, spam content, violent, sarcasm and clickbait all these sources are merged into a single corpus dataset along with their respective labels. The final multiclass corpus merged consists of 126,855 text instances in total out of which 80% (101,484) is taken for training and 20% (25,371) is taken for testing without repetition of instances to avoid data leakage.

3.2 Preprocessing

The preprocessing step is necessary in text classification, particularly with noisy short texts extracted from social media, which involves misleading links and characters that impact the classification performance task. The above-mentioned datasets are generally

unstructured and noisy, so the preprocessing stage is necessary for text classification for preparation and detection, which helps to enhance the model performance. A preprocessing procedure is carried out by using an including the following steps, such as tokenization, stemming, token filtering and stop word removal are as described follows:

Tokenization: A token is a sequence of characters considered as a single unit. The procedure of splitting the text into tokens allows the creation of meaningful features for analysis. In text categorisation, tokens are often chosen from paragraphs, sentences or text segments to denote the content efficiently.

Stemming: Stemming is based on the idea that words with the same stem have similar meanings. By minimizing these words to the common stem NLP model is treated as a single term which helps enhance analysis and processing. The output feature count is reduced when stems are used instead of original words. A stemming model identifies the suffixes in words and trims them to their root form, which is necessary to note that the result is not always be complete or meaningful word.

Token Filtering: The token filtering is performed to eliminate irrelevant or low-quality tokens that do not contribute a meaningful word. This consists of removing punctuation, special characters, numerical values or common words. Through recollecting only informative tokens, the dimensionality of a feature space is minimised and the performance of the text categorisation model is enhanced.

Stop words removal: Stop words are frequently occurring words that do not contribute significantly to the meaning of a sentence. While the model adds little value to the feature matrix, the model is typically eliminated in the preprocessing. However, in some cases, particular stop words still need to be retained for the necessity of preserving the intended meaning. To handle this, a customised dictionary was developed as a substitute of using the NLTK list, which features denotes as a single value feature. If a token of text or sentence involves a stop word from a list, a feature value is set to 1, it is set to 0.

3.3 Bidirectional Encoder Representations Transformer

The BERT-based uncased model denotes a significant enhancement in NLP, generating a robust framework for different types of textual data by its bidirectional transformer architecture. Bidirectional connections within a sequence permit the model to learn deep contextual associations in equally directions within a sequence, which allows deep contextual learning. In a context of violence classification in text, BERT base uncased generates a unique capability for the model is constructed for violence classification in text, enabling to analysis of contextual relationships and the exact recognition of different social media texts like hate speech, cyberbullying, abusive language, racist and sexist comments, fake news, spam content, violence, sarcasm and clickbait. In BERT-based extensive pretraining on large-scale textual data involves clickbait, spam detection, hate speech and abusive language, sarcasm, news headlines, YouTube comments and SMS spam detection datasets. This pretraining permits a model to learn rich linguistic and contextual representations permit it to transfer knowledge efficiently to other classes, such as multiclass classification using text. Through exposure to diverse data during pretraining, BERT enhances a deep understanding of language patterns, improving its ability to generalise effectively and perform well across different text types and classes.

A BERT is operated on a multilayer transformer architecture consisting of 12 transformer encoder layers, where every layer includes dual fundamental mechanisms such as multi-head self-attention and feedforward neural networks. While allowing the model to focus on difficult relationships among tokens, bidirectional processing is difficult for understanding not only the direct sequence of tokens but also the underlying dependencies that show whether a text into the violence class or not.

In a text classification, the capability to focus on relationships among textual generates a significant advantage in accurately recognising a violent class in texts.

Processing starts with tokenization, where the input text is split into subword units by word piece tokenisation. Every token is categorised into a 768-dimensional embedding that captures both semantic and syntactic data, while embeddings are vital for efficiently denoting a feature of multiclass text classification in the path of BERT understanding. A segment embedding is integrated to differentiate among various input sequences, while positional embedding confirms that a BERT model identifies the order of tokens, which is necessary for understanding a temporal connection within text data. A token of transformation is it's based on the embedding vector as presented in Eq. (1):

$$E_t = Embed(t) \quad (1)$$

Once tokenised, the data passes by multiple transformer layers where the multi-head self-attention mechanism allows each token to attend to all other tokens in the sequence while capturing complex dependencies and relationships among them. For text classifications, where relationships among a various class require capturing the self-attention mechanism, which is described in Eq. (2):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

Subsequently, the self-attention procedure an output is passed through an FNN, permitting non-linear transformations of the extracted features. This step refines the input data, permitting the model to learn representations that capture complex and high-level patterns, which are necessary for accurately classifying whether the text into a class or not. A feedforward neural network is expressed in Eq. (3):

$$FFN(x) = ReLU(xW_1 + b_1)W_2 + b_2 \quad (3)$$

A pretraining stage of BERT consists of two key objectives are Masked Language Modelling (MLM) and Next Sentence Prediction (NSP). In an MLM task, certain tokens in the input sequence are masked and the model is trained to predict these missing tokens based on the surrounding context. This allows BERT to learn contextual relationships and dependencies in data, which is necessary for data generalisation across various data is determined in Eq. (4):

$$\mathcal{L}_{MLM} = \sum_{t=1}^T \log P(t|context) \quad (4)$$

In an NSP task, the trained model determines whether one sentence logically follows another by focusing on relationships among adjacent sentences.

In an SDN context, this ability improves BERT-based understanding of sequential text data and the capability to classify irregular classes. The objective function of NSP is mentioned in Eq. (5):

$$\mathcal{L}_{NSP} = -\frac{1}{N} \sum_{i=1}^N [y_i \log P(y_i) + (1 - y_i) \log (1 - P(y_i))] \quad (5)$$

In summary BERT based and pretraining provide an efficient framework for classifying text data into multiple classes. BERT's ability to capture difficult relationships within data makes it prepared to handle the dynamic nature of text data and identify multiple classes through text data classification.

3.3.1. Layer-wise learning rate Decay

During fine-tuning, all layers of BERT are commonly updated with the same learning rate. Nevertheless, the lower transformer layers primarily encode generic syntactic and lexical data, while an upper layer captures task-specific semantics. All layers equally interrupt the pre-trained representations in the lower layers, leading to unstable convergence. To solve this issue, LLRD is presented in this the learning rate to each transformer layer progressively enhances from the high to the low layers. This modification permits the higher layers to capture and learn domain-specific nuances connected to violent or aggressive expressions, while preserving the stable linguistic representations in the lower layers. The BERT consists of various transformer layers while lower layer encodes a semantics and high layer captures a task, operating a single learning rate to all layers during fine tuning disrupt these pretrained representation. The LLRD assigns a distinct learning rates to each transformer layer minimizing progressively from top to bottom. Let L shows the total number of transformer layer in BERT and η is base learning rate assigned in top layer, a learning rate for the l th layer mentioned in Eq. (5):

$$\eta_l = \eta \cdot \lambda^{(L-1)} \quad (5)$$

Where, η_l is the learning rate of layer l , η is maximum learning rate, $\lambda \in (0,1)$ is decay factor. The parameter for rule layer l is presented in Eq. (6):

$$\theta_l^{(t+1)} = \theta_l^t - \eta_l \nabla_{\theta_l} L \quad (6)$$

Where, θ_l is parameters of layer l , t denotes training iteration and L is multiclass classification loss.

The Eq. (6) confirms an small parameter updates in lower and higher layer preserving pretrained linguistic representation while allowing efficient text classification for multiclass violence detection.

3.3.2. Token-Level Class-Wise Importance

Conventional BERT fine-tuning treats every token equally, which leads to suboptimal attention toward critical tokens that show a violent intent. To overcome this challenge, a Token-level Class Importance mechanism is combined into the attention and loss computation layer. During the training stage, every feature is assigned a dynamic weight proportional to its violence or nonviolence text, which is confirmed by a BERT encoder. A token is controlled via a coefficient of class importance using contextual relevance from the final hidden layer output. Through integrating semantic data, a model improves its capability to detect violence when violence cues are contextually masked or appear infrequently. Let the input text sequence denoted in Eq. (7):

$$X = \{x_1, x_2, \dots, x_n\} \quad (7)$$

Where, $x_i \in \mathbb{R}^d$ shows a contextual embedding of i th token attained from a final BERT, y_c is logit based to class c , an importance of token i for class c is mentioned in Eq. (8):

$$I_i^{(c)} = x_i \cdot \frac{\partial y_c}{\partial x_i} \quad (8)$$

Where, $I_i^{(c)}$ is class wise importance score of tokens i , $\frac{\partial y_c}{\partial x_i}$ calculates a sensitivity of class c to token i , to attain a normalized weighted across tokens an scores are calculated as based on Eq. (9):

$$\alpha_i^{(c)} = \frac{|I_i^{(c)}|}{\sum_{j=1}^n |I_j^{(c)}|} \quad (9)$$

Where, $\alpha_i^{(c)}$ is normalized contribution of token i to class c . These weights emphasize semantically meaningful tokens while suppressing irrelevant words which enhancing the model capability to differentiate among normal and violence texts. The attention maps explicitly highlight and prioritise the tokens' influence on violent classification decisions such as from nine classes such as hate speech, cyberbullying, abusive language, racist and sex comments, fake news, spam content, violent, sarcasm and clickbait.

4. Simulation Results and Discussion

A performance of the proposed TLLRD-BERT is evaluated and estimated, which relies on merged multiclass corpus dataset for detection of 9 classes. The software required for the proposed TLLRD-BERT methodology is Python 3.8. Details of the system configuration are provided below.

- Processor: i7
- RAM: 16 GB
- System type: 64-bit OS
- GPU: 6 GB
- Memory: 1 TB

Assessment metrics like accuracy, precision, recall, and F1-score are used to evaluate the TLLRD-BERT. A mathematical formula is given in Eqs. (10)– (13).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (10)$$

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

$$F1 - score = 2X \frac{Precision \times Recall}{Precision + Recall} \quad (13)$$

Where TP - True Positive, TN - True Negative, FP - False Positive, FN - False Negative.

4.1 Performance Analysis

This section presents the qualitative and quantitative results of the proposed TLLRD-BERT, evaluated using by collected dataset to use to estimate the TLLRD-BERT performance. A performance evaluation of TLLRD-BERT is analysed by collecting a dataset and estimating various metrics like precision, recall and F1-score. Table 1 denotes a proposed model performance for the multiclass corpus dataset and various outputs from nine classes, such as hate speech, cyberbullying, abusive language, racist and sex comments, fake news, spam content, violent, sarcasm and clickbait.

Table 1 shows the proposed TLLRD-BERT performance for a class-wise result for a multiclass corpus dataset where different outputs like hate speech, cyberbullying, abusive language, racist and sex comments etc., are attained through precision, recall and F1-

score. A macro avg result of the proposed model achieves a precision of 94.9%, an F1-score of 94.4% and a recall of 94.1%. Table 2 illustrates a classification output of the proposed TLLRD-BERT compared with other classification models.

Table 2 shows that a proposed LTLLRD-BERT model attains an accuracy of 96.8% accuracy, 95.6% of F1-score and a precision of 96.4%. Table 3 illustrates the computational complexity analysis of TLLRD-BERT model for multiclass text classification based on multiclass corpus dataset.

The computational complexity analysis is presented in Table 3 shows that the proposed TLLRD-BERT is significantly effective compared to other methods. While RoBERTa and BERT need high training time, inference time, memory usage and FLOPs because of the huge parameter sizes, Bi-LSTM suffered from high computational overhead caused though sequential processing. The proposed TLLRD-BERT achieves the lowest training and inference time, minimized memory consumption are make the model suitable and computationally efficient in text classification performance. Table 4 ablation study of proposed TLLRD-BERT for multiclass text classification.

Table 4 illustrates the ablation study of performance for BERT, TLCI+BERT, LLRD+BERT and the proposed TLLRD-BERT. The baseline BERT model generates better output which are enhanced via integrating TLCI and LLRD. The proposed TLLRD-BERT integrates both mechanisms permitting efficient contextual feature learning and stable fine-tuning across layers.

This combined model increases representation quality and generalization allowing the proposed TLLRD-BERT to achieve the high level of values compared to all methods. Table 5 illustrates the statistical analysis of proposed TLLRD-BERT for text classification.

The Table 5 depicts the statistical significance and stability analysis of text classification models by p-values from the t-test and the SD of accuracy. Bi-LSTM and RoBERTa exhibit higher p-values and larger standard deviations shows the high level of performance variability. BERT shows improved stability with a lower p-value and minimized deviation. The proposed TLLRD-BERT attains the lowest p-value of 0.04 and the SD of ± 1.28 determines the statistically significant which ensuring the robustness of the proposed TLLRD-BERT approach.

Table 1. The class-wise classification performance of the proposed TLLRD-BERT model on the multiclass corpus dataset

Classes	Precision (%)	Recall (%)	F1-score (%)
Hate speech	94.2	93.6	93.9
Cyberbullying	95.1	94.3	94.7
Abusive language	93.7	92.4	93.1
Racist and sex comments	94.8	93.5	94.1
Fake news	96.3	95.5	95.9
Spam content	97.0	96.1	96.5
Violent	95.8	94.9	95.3
Sarcasm	92.6	91.8	92.2
Clickbait	94.5	93.7	94.1
Macro Avg	94.9	94.1	94.4

Table 2. Performance analysis of the proposed TLLRD-BERT method with existing approaches for classification in multiclass corpus dataset

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Bi-LSTM	91.2	90.8	89.7	90.2

RoBERTa	93.8	93.1	92.5	92.8
BERT	94.6	94.1	93.3	93.6
Proposed TLLRD-BERT	96.8	96.4	95.3	95.6

Table 3. performance analysis of computational complexity of proposed TLLRD-BERT for multiclass text classification

Methods	Training Time (s)	Inference Time (s)	Memory Usage (KB)	FLOPS (G)	Parameters (s)
Bi-LSTM	185	0.082	2420	7.6	102
RoBERTa	520	0.210	3680	11.8	152
BERT	460	0.185	3420	9.6	110
Proposed TLLRD-BERT	140	0.058	1420	6.1	95

Table 4. Ablation study of BERT-based models for text classification using accuracy, precision, recall, and F1-score

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
BERT	94.6	94.1	93.3	93.6
TLCI+BERT	95.4	95.6	94.2	94.6
LLRD+BERT	96.1	95.8	94.8	95.2
Proposed TLLRD-BERT	96.8	96.4	95.3	95.6

Table 5. Statistical analysis of proposed TLLRD-BERT for text classification based on multi corpus dataset

Methods	p-value from t-test	Standard Deviation (SD) from accuracy value
Bi-LSTM	0.11	± 1.80
RoBERTa	0.09	± 1.43
BERT	0.07	± 1.38
Proposed TLLRD-BERT	0.04	± 1.28

This improvement is achieved through a layer-level representation integration and refined contextual embedding alignment in TLLRD-BERT, which allows a model to capture both shallow and deep semantic features efficiently while increasing the capture of subtle contextual and minimising false positives. The proposed TLLRD-BERT reduced information loss across layers and increased discriminative feature learning output in increased detection of difficult linguistic patterns and model classification performance.

4.2 Analysis of the Confusion Matrix

This section presents a confusion matrix of three datasets, such as news headlines, YouTube comments and SMS spam detection datasets. Fig. 2-4 depicts the confusion matrix of news headlines, YouTube comments and SMS spam detection datasets model classification performance. A confusion matrix in Fig 2.

2-4 illustrates the classification capability of the proposed TLLRD-BERT model across three benchmark datasets. Every matrix visually denotes a model's effectiveness in reducing false predictions and maximising true class recognition.

In Fig. 2, a news headline dataset classifies true positive samples and false samples as fake with few misclassifications. A high concentration along with the region confirms a

strong model capability in detecting fake vs real news headlines. A minor classification is primarily because of overlapping patterns among satirical and misleading headlines.

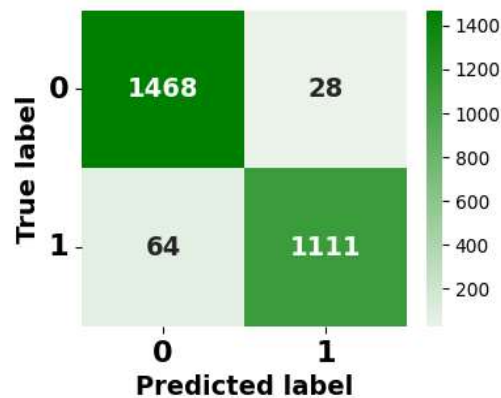


Figure. 2 The confusion matrix of the news headline dataset illustrates an accurate classification of real and fake news using the proposed TLLRD-BERT

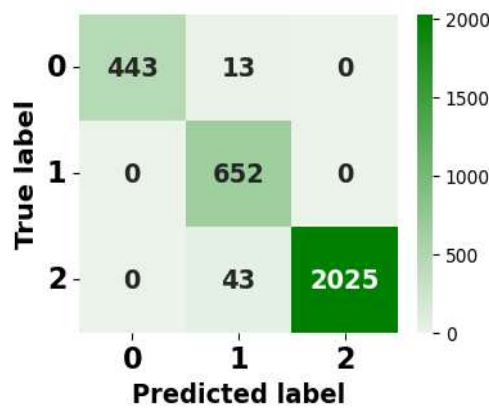


Figure. 3 Confusion matrix of the YouTube comments dataset demonstrates an precise classification using TLLRD-BERT

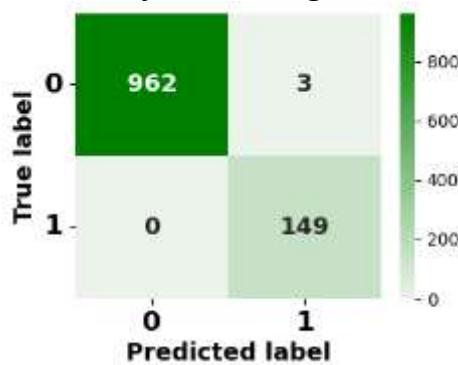


Figure. 4 The confusion matrix of an SMS spam detection dataset shows a differentiation among spam and ham with minimal misclassification

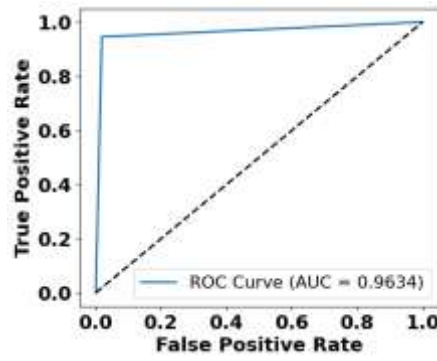


Figure. 5 ROC curve of the News headline dataset representing enhanced classification performance with higher AUC values

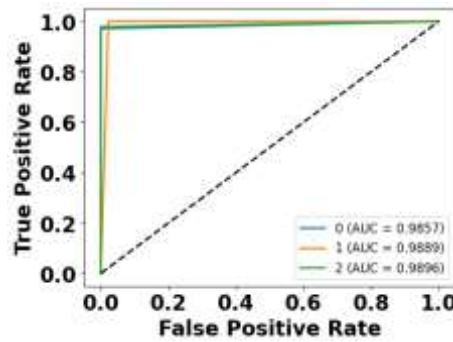


Figure. 6 ROC curve for youtube comments attaining a higher AUC value in the proposed TLLRD-BERT

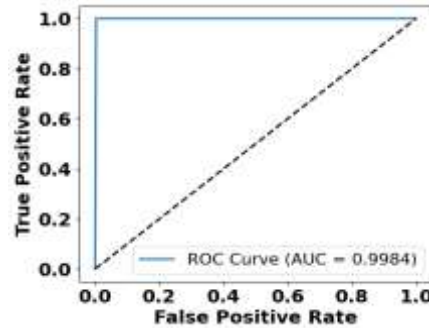


Figure. 7 ROC curve of the SMS spam detection dataset shows a better performance between spam and ham in the proposed TLLRD-BERT

In Fig. 3, the YouTube comments dataset is split into positive, negative and neutral comments. The high true positive indicates the proposed model's capability to accurately handle informal comments commonly found in social media comments. A minimal false negative rate depicts the effectiveness of the proposed TLLRD-BERT effectiveness of layer-level representation association in capturing contextual subtle.

Fig. 4 presents the SMS spam detection dataset confusion matrix reports that ham or spam in that refined a contextual embedding in TLLRD-BERT in recognizing a hidden spam when differentiating as spam or ham, which contributes to enhancing a model's performance.

4.3 Analysis of ROC Curve

This section demonstrates the ROC curves of the proposed TLLRD-BERT model, shows the classification performance across all classes in three different datasets. The curve

indicates strong class separability and high discrimination capability. Figs. 5-7 shows the ROC curve of news headlines, YouTube comments and SMS spam detection datasets, respectively.

Figs. 5-7 illustrate the ROC curves, showing the better classification and generalization capability of the proposed TLLRD-BERT across different datasets. Fig. 5 demonstrates that the model efficiently differentiates texts among a news headline dataset, which highlights its ability to capture contextual information in textual data. Fig. 6 depicts the performance of the proposed model on the YouTube comments dataset, while the proposed TLLRD-BERT attains a high AUC value across all classes, ensuring stability and generalisation. At last, Fig. 7 presents that highlighting a model accuracy and reliability in recognising true positives while reducing false positive classification.

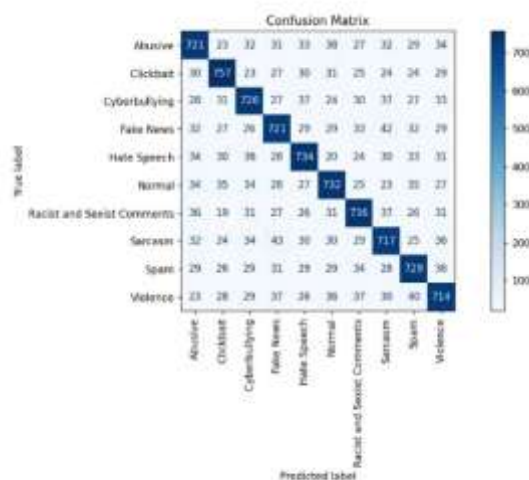


Figure 8 illustrates the unified dataset across all nine categories, emphasizing the classification outcomes of true positive and false samples as harmful.

In Figure 8, the unified dataset containing all nine categories demonstrates the classification of true positive and false samples as harmful, with a few misclassifications observed. The Y-axis of the figure represents the actual categories, while the X-axis denotes the predicted classes generated by our classifier. The numerical values indicate the count per 1,000 samples. For instance, Abusive vs Abusive = 721 implies that 721 out of 1,000 abusive samples were correctly classified, whereas Abusive vs Clickbait = 23 indicates that 23 abusive samples were incorrectly classified as clickbait. Although the proposed TLLRD-BERT model shows a minor compromise in accuracy, it effectively distinguishes whether the given content is harmful or not.

4.4 Comparative Analysis

The proposed TLLRD-BERT is compared with various datasets such as the news headline dataset, the YouTube comments dataset and the SMS spam detection dataset to perform in comparative analysis.

Table 6. Comparative analysis of the YouTube comments dataset is compared with existing approaches in the proposed TLLRD-BERT

Methods	Accuracy (%)
FCNN [14]	98
CBLSTM [16]	95.21
Proposed TLLRD-BERT	98.24

Table 7. Comparative analysis of proposed TLLRD-BERT using the News headline dataset for proposed classification

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
voting based ensemble model [18]	94.89	94.25	94.95	94.60
CNN+ Bi-LSTM [20]	85.32	84.56	83.42	83.39
Proposed TLLRD-BERT	96.54	96.78	95.17	97.45

Table 8. Comparative analysis of SMS spam detection dataset results for the proposed TLLRD-BERT model for SMS spam text classification

Methods	Accuracy (%)
CBLSTM [16]	99.30
LSTM+CNN [19]	99.56
Proposed TLLRD-BERT	99.71

The proposed TLLRD-BERT model is evaluated by existing models like FCNN [14], CBLSTM [16], LSTM+CNN [19] and CNN+ Bi-LSTM [20]. Tables 6, 7 and 8 show a comparative analysis of YouTube comments, news headlines and SMS spam detection datasets compared with the proposed TLLRD-BERT.

Tables 6-8 demonstrate that a proposed TLLRD-BERT consistently outperforms an existing approach across all datasets, such as YouTube comments, news headlines and SMS spam detection datasets. A model achieves a higher level of accuracy, precision, recall and F1-score value because of its transformer-guided layer-level representation and a refined contextual embedding mechanism, which allows a deeper semantic understanding in text and splits the classes. By capturing both local and global dependencies, the proposed TLLRD-BERT model reduces misclassifications and improves its ability to address discrimination challenges in textual patterns, resulting in enhanced classification accuracy across different contexts. The proposed TLLRD-BERT attains an accuracy of 98.24%, 96.54% and 99.71% through the YouTube comments dataset, the news headline dataset and the SMS spam detection dataset, respectively.

4.5 Discussion

In this section denotes the advantages of proposed TLLRD-BERT and limitations of the existing approaches are discussed in this section. The limitations of existing approaches such as preprocessing, hand crafted features and hybrid architectures are limited the model generalization and scalability. The existing models are difficult with contextual ambiguity, sparse semantic cues specifically in sarcasm and abusive content detection. The proposed TLLRD-BERT is overcome these challenges, LLRD is presented in this the learning rate to each transformer layer progressively enhances from the high to the low layers. Token-level Class Importance mechanism is combined into the attention and loss computation layer. During the training stage, every feature is assigned a dynamic weight proportional to its violence or nonviolence text which is confirmed by a BERT encoder.

5. Conclusion

In this research, the proposed TLLRD-BERT is used to enhance text classification and is split into nine classes, such as hate speech, cyberbullying, abusive language, racist and sex comments, fake news, spam content, violent, sarcasm and clickbait, which are obtained by the collected dataset. An TLCl and LWLRD are integrated through progressively adjusting a learning rate across layers and by dynamically modifying the weight assigned to every important token; the model obtains enhanced training stability. In the preprocessing stage, methods such as tokenization, stemming, token filtering and stop word removal are employed to recollect only meaningful terms and standardise the text representation. The proposed TLLRD-BERT attains an accuracy of 98.24%, 96.54% and 99.71% through the YouTube comments dataset, the news headline dataset and the SMS spam detection dataset, respectively. The future work will focus on extending the transformer framework for multilingual datasets and incorporating lightweight transformers for real-time deployment.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Pooja Pandey contributed to the conceptualization, methodology, analysis, and implementation of the study. Dr. Sharvari Govilkar contributed to validation and proofreading. Both authors reviewed and approved the final version of the manuscript.

References

- [1] L. Ou, and Z. Li, "Modeling inter-modal incongruous sentiment expressions for multi modal sarcasm detection", *Neurocomputing*, Vol. 616, p. 128874, 2025.
- [2] S. Tuarob, M. Satravisut, P. Sangtunchai, S. Nunthavanich, and T. Noraset, "FALCoN: Detecting and classifying abusive language in social networks using context features and unlabeled data", *Information Processing & Management*, Vol. 60, No. 4, p. 103381, 2023.
- [3] F.Z. El-Alami, S.O. El Alaoui, and N.E. Nahnahi, "A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model," *Journal of King Saud University-Computer and Information Sciences*, Vol. 34, No. 8, pp. 6048-6056, 2022.
- [4] B.R. Chakravarthi, M.B. Jagadeeshan, V. Palanikumar, and R. Priyadharshini, "Offensive language identification in dravidian languages using mpnet and cnn", *International Journal of Information Management Data Insights*, Vol. 3, No. 1, p. 100151, 2023.
- [5] K.B. Nelatoori and H.B. Kommanti, "Attention-based bi-lstm network for abusive language detection," *IETE Journal of Research*, Vol. 69, No. 11, pp. 7884-7892, 2022.
- [6] X. Wang, Y. Wang, D. He, Z. Yu, Y. Li, L. Wang, J. Dang, and D. Jin, "Elevating knowledge-enhanced entity and relationship understanding for sarcasm detection", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 37, No. 6, pp. 3356-3371, 2025.
- [7] S. Khan, I. Qasim, W. Khan, K. Aurangzeb, J. A. Khan, and M. S. Anwar, "A novel transformer attention-based approach for sarcasm detection", *Expert Systems*, Vol. 42, No. 1, p. e13686, 2025.
- [8] Y.W. Ma, J.L. Chen, L.D. Chen, and Y.M. Huang, "Intelligent clickbait news detection system based on artificial intelligence and feature engineering", *IEEE Transactions on Engineering Management*, Vol. 71, pp. 12509-12518, 2022.
- [9] T.T. Aurpa, R. Sadik, and M.S. Ahmed, "Abusive Bangla comments detection on Facebook using transformer-based deep learning models", *Social Network Analysis and Mining*, Vol. 12, No. 1, p. 24, 2021.

- [10] O. Galal, A.H. Abdel-Gawad, and M. Farouk, "Rethinking of BERT sentence embedding for text classification", *Neural Computing and Applications*, Vol. 36, No. 32, pp. 20245-20258, 2024.
- [11] S.M.A.H. Shah, S.F.H. Shah, A. Ullah, A. Rizwan, G. Atteia, and M. Alabdulhafith, "Arabic sentiment analysis and sarcasm detection using probabilistic projections-based variational switch transformer", *IEEE Access*, Vol. 11, pp. 67865-67881, 2023.
- [12] R. Fatima, M.M.S. Fareed, S. Ullah, G. Ahmad, and S. Mahmood, "An optimized approach for detection and classification of spam email's using ensemble methods", *Wireless Personal Communications*, Vol. 139, No. 1, pp. 347-373, 2024.
- [13] J. Bacha, F. Ullah, J. Khan, A.W. Sardar, and S. Lee, "A deep learning-based framework for offensive text detection in unstructured data for heterogeneous social media", *IEEE Access*, Vol. 11, pp. 124484-124498, 2023.
- [14] M. Anand, K.B. Sahay, M.A. Ahmed, D. Sultan, R.R. Chandan, and B. Singh, "Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques", *Theoretical Computer Science*, Vol. 943, pp. 203-218, 2023.
- [15] T. Yue, R. Mao, H. Wang, Z. Hu, and E. Cambria, "KnowleNet: Knowledge fusion network for multimodal sarcasm detection", *Information Fusion*, Vol. 100, p. 101921, 2023.
- [16] R. Ghanem and H. Erbay, "Spam detection on social networks using deep contextualized word representation", *Multimedia Tools and Applications*, Vol. 82, No. 3, pp. 3697-3712, 2022.
- [17] D. Vinoth and P. Prabhavathy, "An intelligent machine learning-based sarcasm detection and classification model on social networks", *The Journal of Supercomputing*, Vol. 78, No. 8, pp. 10575-10594, 2022.
- [18] R.T. Gedela, U. Baruah, and B. Soni, "Deep contextualised text representation and learning for sarcasm detection", *Arabian Journal for Science and Engineering*, Vol. 49, No. 3, pp. 3719-3734, 2024.
- [19] H.A. Al-Kabbi, M.R. Feizi-Derakhshi, and S. Pashazadeh, "Multi-type feature extraction and early fusion framework for sms spam detection", *IEEE Access*, Vol. 11, pp. 123756-123765, 2023.
- [20] C. Thaokar, J.K. Rout, M. Rout, and N.K. Ray, "N-Gram based sarcasm detection for news and social media text using hybrid deep learning models", *SN Computer Science*, Vol. 5, No. 1, p. 163, 2024.
- [21] K. Arunthavachelvan, S. Raza, C. Ding, and E. Hasan, "PL-NCC: a novel approach for fake news detection through data augmentation", *Social Network Analysis and Mining*, Vol. 14, No. 1, p. 217, 2024.
- [22] E.Y. Daraghmi, S. Qadan, Y.A. Daraghmi, R. Yousuf, O. Cheikhrouhou, and M. Baz, "From text to insight: An integrated cnn-bilstm-gru model for arabic cyberbullying detection", *IEEE Access*, Vol. 12, pp. 103504-103519, 2024.
- [23] Ç.O. Aliyeva and M. Yağanoğlu, "Deep learning approach to detect cyberbullying on twitter", *Multimedia Tools and Applications*, Vol. 84, No. 19, pp. 20497-20520, 2024.
- [24] Clickbait dataset access link:
<https://www.kaggle.com/datasets/amananandrai/clickbait-dataset/code> (Accessed on October 2025).
- [25] Hate speech and abusive language dataset access link:
<https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset> (Accessed on October 2025).
- [26] Sarcasm detection dataset access link:
<https://www.kaggle.com/datasets/saurabhbhagchi/sarcasm-detection-through-nlp> (Accessed on October 2025).

- [27] News headline dataset access link: <https://www.kaggle.com/datasets/rmisra/news-headlines-dataset-for-sarcasm-detection> (Accessed on October 2025).
- [28] YouTube comments dataset access link:
<https://www.kaggle.com/datasets/atifaliak/youtube-comments-dataset> (Accessed on October 2025).
- [29] SMS spam detection dataset access link:
<https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset> (Accessed on October 2025).