Kufa Journal of Engineering
Vol. 17, No. 2, April 2026, P.P. 79-100
Article history: Received 12 January 2025, last revised 6 March 2025, accepted 21 March 2025

OPTIMIZING SENTIMENT ANALYSIS WITH BERT ENHANCED BY BILSTM AND BIGRU LAYERS

V. Uma[1] and V. Ganesh[2]

[1] Department of Computer and Information Science, Annamalai University, Cuddalore 608002, Tamil Nadu, India, Email: uma72sekar65@gmail.com.

[2] Department of Computer Science, Government Arts College Autonomous, Kumbakonam 612002, Tamil Nadu, India, Email: dinesh59@gmail.com.

https://doi.org/10.30572/2018/KJE/170206

ABSTRACT

As mobile technology develops rapidly, social media becomes a rich source of platform for people to voice for his or her opinions and point of views. In order to support business decision making and policy making, we need to quantify public moods, sentiments by conducting sentiment analysis which emerges as a key research area in recent year. Many research efforts have been put into performing sentiment analysis, many tools and algorithms are devised to determine the sentiment as positive, negative or neutral, in social media. We improve the sentiment classfier by training a classical machine learning model on three different data sets. Although BERT has shown good performance in sentiment analysis, we need to find a way to raise the accuracy. We propose four Deep Learning models by combining BERT with Bidirectional Long ShortTerm Memory (BiLSTM) and Bidirectional Gated Recurrent Unit (BiGRU) algorithms. Our paper uses pre, trained word embedding vectors as inputs and aims to increase the accuracy of sentiment analysis and test the effect of BERT‘s combination with BiGRU and BiLSTM layers, DistilBERT and RoBERTa. We test the classification of text sentiments with and without use of emoji symbols. A comparative analysis of 2 pre, trained BERT models and 7 classical machine learning models suggests that the models with BiGRU layers achieve best performance in our sentiment analysis pipelines.

KEYWORDS

Sentimental analysis, Machine learning, LSTM, Attention mechanism.

1. INTRODUCTION

Amid the perpetually altering landscape of the information age, the pervasive presence of mobile technology has inaugurated a novel age of communication in which the social media platforms are transformed into today's agora, the public space where citizens express their thoughts, beliefs, and feelings (Abdul-Mageed, Ungar, et al., 2017; Almatrafi, Parack, et al., 2015; Das, Gamback, et al., 2012; Giatsoglou, et al., 2017; He, et al., 2012). The global online space serves as an exceptional space for people to convey their feelings on countless subjects, which range from common daily events to rather serious and significant topics (Karyotis, et al., 2018; Maas, et al., 2011; Njølstad, et al., 2014). Information acquired from the aggregate emotion of online communities are irreplaceable, not only for businesses struggling in a competitive marketplace, but also for political entities making knowledgeable choices.

One of the most cutting-edge uses of artificial intelligence and natural language processing, sentiment analysis is indispensable in exploring the nuances of public opinion. Through identifying the polarity (positive, negative or neutral) of a sentiment, it has enabled us to explore the fluctuating emotions of this technological age (Pang, Lee, et al. 2002, Xia, Zong, et al. 2011). This information saturated time sees sentiment analysis acting as a compass to navigate the vast sea of social media data.

This work is a reaction to the growing interest in sentiment analysis as a consequence of digital communication expansion. In the given study, the central goal is the increased accuracy in sentiment analysis due to the fact that there is a lot of ambiguity and context which ordinary approaches lack (Nimmi, et al., 2022; Adoma, et al., 2020; Sirisha & Bolem, 2022; Bansal & Srivastava, 2019; Ma, et al., 2018; Zainuddin, et al., 2018). Although the state of the art BERT (Bidirectional Encoder Representations from Transformers) model has obtained decent performance on sentiment analysis, there is still space for improvements (Prottasha, et al., 2022; Wang, et al., 2013; Demotte, et al.,2023 ; Janjua, et al., 2021; Thapa, 2022) and hence the urge for better accuracy of the systems has motivated us to propose an effective method combined with BERT and Bidirectional Long Short-Term Memory (BiLSTM) and Bidirectional Gated Recurrent Unit (BiGRU) algorithms (Indrayuni & Nurhadi, 2020; Dang, et al., 2020; Kumawat, et al., 2021; Xiang, et al., 2020; Wen & Li, 2018; Janjua, et al., 2021; Tan, et al., 2022; Mohammed, et al., 2022; Hashim & Mazinani, 2025).

We utilize pre-trained word embedding vectors as base for fine-tuning these hybrid models, since these vectors provides basic level of semantic understanding and contextual awareness for the human sentiments. In our study, our research would contribute in two ways, which not only aiming to improve the sentiment analysis but also to study the effect of embedding the

BiGRU/BiLSTM layers within the BERT-based models (such as DistilBERT, RoBERTa). It includes sentiment analysis tasks on texts both with and without emojis.

In order to evaluate our proposed approaches, we carried out a thorough comparison with two pre-trained BERT architectures and several (seven) classical machine learning approaches. As explained in the abstract, we find that the proposed BiGRU incorporated architecture provides promising improvements over two pre-trained BERT models, and a range of ML based approaches.

What is more important in the proposed work, we analyze the problem posed by the short text form of tweet where the short text plays a crucial role in classification. Even though BERT has already been proved effective in analyzing tweet sentiment but more accuracy is needed. For better representation of short text, we propose a novel model which combines the idea of BERT and BiLSTM and BiGRU layers. The BERT model is used to learn word vectors with training, which makes short text word representation better, and the Bidirectional LSTM (BiLSTM) and Bidirectional GRU (BiGRU) networks are used to model and learn the characteristics of sequences. We also will investigate the effect of number of layers and their positions of BiLSTM/BiGRU on model performance.

The principal contributions of this research can be encapsulated as follows:

Four novel hybrid deep learning models which were created for the specific task of emotion classification across three different datasets were introduced. These models are made up of four models that use RoBERTa, and four that use DistilBERT, these models are then critically assessed to find which hybrid model is the most competent at obtaining the context from a given text input.

By using BiGRU or BiLSTM networks to capture the useful context from the text to improve the fine-tuning process.

A detailed investigation on the impact of emojis as predictive features in the classification models. This may include training models with emojis included in the training dataset and performing experiments to examine the change in results without emojis in the pre-processing stage.

2. RELATED WORK

Many researchers have made many contributions to the field of sentiment analysis and emotion recognition using modern natural language processing methods and machine learning models. Research in these fields covered many areas of sentiment analysis, emotion recognition and

aspect based sentiment analysis. The following section briefly describes some of these research studies

In Nimmi, et al. (2022), AVEDL (Average Voting Ensemble Deep Learning) Model analyzed calls to emergency response assistance system (ERSS) when COVID-19 was happening. The AVEDL Model used a pre-trained transformer-based models including BERT, DistilBERT and RoBERTa for the purpose of eliciting emotions and sentiment from the emergency calls. It presented an accuracy of 86.46% and a Macro-average F1-score of 85.20%. Its weaknesses include a poor recognition of slang, different speech patterns used, which often happens during emergencies, thus difficulty in calculating COVID-19 related emotions.

The research described in Adoma (2020) focuses on evaluating pre-trained transformer models namely BERT, RoBERTa, DistilBERT and XLNet to detect emotions in text. These pre-trained models were fine-tuned on the ISEAR data to identify emotions, and these results were significant. Of the transformer models considered, RoBERTa gave an accuracy of 74%. However, although these results are important for emotion detection, generalization for new languages is not demonstrated here.

In Sirisha and Bolem (2022), the aim was to carry out aspect- based sentiment analysis with RoBERTa and LSTM for Twitter data of Ukraine Conflict. The research was successfully conducted in exploring emotions like Optimism, Sadness, Anger and Joy with accuracy of 94.7% as achieved which is remarkable. There is limitation in this research being domain-specific (Twitter data of a particular event) and in extending it to several domains and data.

Bansal and Srivastava (2019) proposed an attribute-based hybrid approach for customer intelligence analysis. They employed POS tags for identifying the attributes. The work mainly focuses on identifying subjects with all the attributes and lowering down the cost of computation. But one drawback is that this may be applied to the short text categorization, the human labeled attribute related words in the dictionaries need to be further researched.

Ma, et al. (2018) built a classification model of LSTM to detect aspect information using common-sense knowledge. The Sentic LSTM model incorporated a recurrent additive network with LSTM and achieved an F-measure of 75.51%, which is acceptable. However, controlling some aspects of hybrid association rule mining is difficult using this approach when used in real-world applications with real-world data.

In Zainuddin, Selamat, et al. (2018), authors used SVM combined with Principal Component Analysis (PCA) andPOS tagsas features to extract for the sentiment classifier. High accuracy values for STS and STC datasets were obtained. The accuracies achieved are higher than most established baseline sentiment classification models. But their solution would be difficult to

use for other social media data such as YouTube and Facebook. In this paper the accuracy ofBangla-BERTwith LSTMs was very high (94.15%). The authors recommended using higher deep learning algorithms with more varied, richer, more balanced data.

Prottasha, Sami, et al. (2022) had implicitly derived attributes by applying a hybrid association rule mining and five utilization methods were proposed. However, this technique had mainly re-captured the explicit ones as it had not handled the implicit words due to context constraints. Although the F-measure of 75.51% was achieved, this technique has a few disadvantages, such as a few factors have difficult to be controlled in practical life of hybrid association rule mining process.

Some prominent limitations across the literature can be observed. Many of the approaches that have been suggested, and that many models are built upon, are inherently domain-specific, which is an undesirable characteristic that limits their generalizability to more general scenarios and data. A prominent and on-going problem, which many sentiment analysis models still find very difficult, is correctly classifying instances of colloquial language and varied speech patterns. Also, many of the aforementioned approaches target a particular language, thus the issue of multi-language analysis has proven to be challenging. Bias within the training data is also a serious limitation to the accuracy of any sentiment analysis model. There may be other practical limitations such as computational resources required. It is also important for models to be interpretable and usable for the end-user. The approaches proposed have some positive contributions to sentiment analysis and emotion recognition but are ultimately limited by issues such as the afore mentioned.

3. METHODOLOGY

3.1. Data Collection

The experiments were started by gathering wide range and variety of data for the sentiment analysis (with emphasis on tweets) in order to encompass a broad range of sentiments and emotions present in short text data. The datasets utilized were:

Airlines Dataset: This dataset consisted of 14,640 tweets classified in different sentiment classes, and specifically, had 2,363 positive tweets, 9,178 negative tweets and 3,099 neutral tweets.

CrowdFlower Dataset: With a total of 3804 tweets, the crowdflower dataset has provided an alternative approach to sentiment analysis. The dataset has a total of 423 post, 1219 negative and 1219 neutral tweets.

Apple Dataset This is another relatively small dataset which has been included for its specific

properties. This dataset has 1,630 tweets. Of the tweets, 686 were post tweets, 143 negative and 801 neutral.

3.2. Preprocessing

Preprocessing is a critical step in sentiment analysis as it directly influences the quality and accuracy of the results. The preprocessing pipeline in this study follows multiple steps, ensuring the refinement of textual data before feature extraction and model training. The primary preprocessing steps include:

1. **Tokenization**: Tokenization is a process where the text is divided into single words or terms which then can be further processed or analyzed. It is one of the key tasks of natural language processing (NLP) which processes the unstructured text and prepares it into a suitable format for use with machine learning. Existing studies by Abdul-Mageed & Ungar (2017) and Maas et al. (2011) has discussed the significant role of tokenization for machine learning based sentiment analysis by dividing text into pieces and analysing them. Standard tokenization is applied using NLTK and spaCy tool kits which enables dealing with word boundaries. The division ensures that a long complicated sentence is not dealt with as a whole and its overall meaning remains intact.

2. **Stopword Removal**: Stopwords (e.g., "the," "is," "in") are removed to reduce noise in textual data (Pang, Lee, & Vaithyanathan, 2002; Bansal & Srivastava, 2019). Stopwords are informative words which appear frequently in the dataset; they eliminate the unnecessary computations and improve the prediction of sentiment classification. This elimination of stopwords ensures robustness for large scale sentiment classifiers, as experimented by Prottasha et al. (2022) and Kumawat et al. (2021). We remove stopwords based on pre-defined stopwords list from NLTK and custom built stopword dictionary for specific domain of the dataset.

3. **Stemming and Lemmatization**: Stemming reduces words to their stem or root forms, whereas Lemmatization is the process of converting words to their dictionary base forms. Stemming and Lemmatization reduce feature space dimensionality and are used for normalizing words. Porter Stemmer and WordNet Lemmatizer are widely used as discussed in Xia et al. (2011) and Demotte et al. (2023). Stemming is fast and effective whereas Lemmatization is more precise since it takes into account the meaning of the words. This is observed by Indrayuni and Nurhadi (2020) and Janjua et al. (2021). Truncating words using stemming leads to a loss of their meanings sometimes, but in Lemmatization the meaning of the word remains correct from the linguistic point of view. This research mainly uses Lemmatization with advanced lemmatizer in spaCy, so that we can obtain a higher quality word representation for our sentiment analysis.

4. **Part-of-Speech (POS) Tagging**: Stemming reduces words to their stem or root forms, whereas Lemmatization is the process of converting words to their dictionary base forms. Stemming and Lemmatization reduce feature space dimensionality and are used for normalizing words. Porter Stemmer and WordNet Lemmatizer are widely used as discussed in Xia et al. (2011) and Demotte et al. (2023). Stemming is fast and effective whereas Lemmatization is more precise since it takes into account the meaning of the words. This is observed by Indrayuni and Nurhadi (2020) and Janjua et al. (2021). Truncating words using stemming leads to a loss of their meanings sometimes, but in Lemmatization the meaning of the word remains correct from the linguistic point of view. This research mainly uses Lemmatization with advanced lemmatizer in spaCy, so that we can obtain a higher quality word representation for our sentiment analysis.

5. **Normalization**: Text Normalization helps resolve differences in spelling, case and abbreviations. Text normalization process standardizes unstructured text to a single format. Several research works highlight the significance of this step-in feature extraction for textual data (Das & Gamback, 2012, Adoma, 2020). Lowercasing the text, dealing with the contractions (for example 'isn't' becomes 'is not') and removing trailing and leading spaces, reduces variance in the text (Zainuddin et al., 2018, Dang et al., 2020). However, in this study, the domain specific text normalization was applied, such that, sector-specific terminology and emotion words were retained but the text was normalized for feature extraction.

6. **Noise Removal**: Filtering unnecessary elements including URLs, special characters, hashtags, repeated letters (Karyotis et al., 2018; Wang et al., 2013) will produce text that has less noise. We used regex-based filtering and NLP based transformation methods to obtain cleaner data for the task. The fact that noise filtering enhances accuracy for sentiment classification and avoids misleading correlations was observed (Nimmi et al., 2022; Xiang et al., 2020). Also, unlike previous approaches, we address emojis by discarding them or by converting them into textual format with the help of emoji dictionaries.

3.3. Definition and Removal of Noise

Noise in textual data refers to irrelevant or redundant information that does not contribute to sentiment classification. It includes:

- **Punctuation and Special Characters**: Using symbols such as ”@” and ”#” and also emojis increase the complexity. Previous researches (Njølstad et al., 2014; Indrayuni & Nurhadi, 2020) prove their removal increases the accuracy. For instance, emoticons may cause problems in sentiment classification, unless processed efficiently. In this research, all emoticons are

systematically replaced with a similar word (e. g. :), > happy) using an emoji lexicon, and this makes the sentiment signal cleaner.

- **URLs and User Mentions**: URLs and mention (e. g. @user) can be often found in social media posts. Related work (Almatrafi et al., 2015; Ma et al., 2018) about the influence of those two entities on text classification and sentiment analysis indicated that the urls must be removed from the sentence in order to avoid any class propagationbias. URLs should be properly handled and instead of @user is chosen the generic term @mention, to maintain sentence syntax and not create a new feature.
- **Stopwords**: Remove the meaningless, widely used words i. e. stopwords (Bansal & Srivastava, 2019; Kumawat et al., 2021). Stopwords are sentiment irrelevant and so it is expected to increase the performance of the sentiment classification task. We applied domain adaptive stopword filtering technique, which retains the sentiment carrying words in the corpus, but discards general, non, carrying words.
- **Repeated Letters and Words**: In an informal text there is sometimes character extension (e. g "goooood"). Normalisation techniques, which fix these words, include character elongation reduction (He, 2012; Prottasha et al., 2022). This is where the word elongation is replaced by the dictionary definition of the word. This research implements a heuristic based method of reducing the elongated words by replacing them by their dictionary definition (e. g "soooo happy" is replaced with "so happy").

We used regex- based filtering along with NLP libraries (Nimmi et al., 2022; Xiang et al., 2020) to remove noise to feed to the next layer of sentiment analysis. Some sophisticated text cleaning techniques like text preprocessing pipeline on tensorflow, pytorch NLP libraries are also explored to clean up the text before model training (Demotte et al., 2023; Thapa, 2022). We also wrote customized scripts for the dataset specific noise removals.

3.4. Model Architecture

A sophisticated and very detailed hybrid deep learning model, as well as comparison against the traditional ML models served as the back bone of this research methodology. It was aimed at analyzing the nitty, gritty aspects of a sentiment analysis of short texts specifically for tweets. In this part the proposed architecture is described in details:

3.4.1. Hybrid Deep Learning Model:

The deep learning model used was a combination of the following indispensable neural network components, using the advantages of BERT, BiLSTM and BiGRU: BERT (Bidirectional Encoder Representations from Transformers) At the heart of the model lies BERT, the building

block of the model. BERT was fundamental in building the capabilities of the model as it provided a way for training contextual word vectors which improved the text representation of short text. BERT was trained in a manner so as to understand context bidirectionally, and using pre-trained embeddings it formed a robust language processing unit. BiLSTM (Bidirectional Long Short-Term Memory) In addition to BERT, the incorporation of BiLSTM contributed to enhance the feature extraction of sentence sequence attributes from text. BiLSTM leverages its bidirectional capability to better capture contextual information, bringing more context and dependencies in textual sequences. This thus equips the model to gain a higher level of sophistication in terms of sentence sequence analysis. BiGRU (Bidirectional Gated Recurrent Unit) Just like BiLSTM, BiGRU was also incorporated to gain the feature extraction capability of sentence sequence attributes. It also complements BiLSTM by offering similar features of bi-directionality to learn text features. Fig.1 shows the deep learning model for processing tweet data. Raw text inputs were mapped to dense numerical vector form using the Embedding layer, and a combination of embeddings (word, position, segment) were utilized in order to encode the semantic meaning. Embeddings were processed by RoBERTa, BiLSTM, and Attention Mechanism that enabled capturing of the dependencies between tweet sequence data and important features that will help in improving the representation of features by allowing the model to focus more on prominent input words and at the same time neglecting the less pertinent input information and so also finding key sentiment indicators in the tweet. Eventually these features were combined, passed through the dense and softmax layers of the Classification layer for predicting the respective sentiments: positive, negative and neutral.

3.4.2. Attention Mechanism

Attention Mechanism benefits sentiment classification task as it enables to give various importance to each of the individual words in the sequence being used which is a very common issue. As majority of the revenue for sentiment analysis for a short text is dependent on few small pieces in context that related with them. Attention layer can effectively pay attention on them and get away the others. In this work, attention mechanism is applied following BiLSTM and BiGRU, which calculatethe weights value for each word. The attention scores are computed as:

$$\alpha_i = \frac{\exp(e_i)}{\sum_j (e_j)}$$

where e_i represents the computed importance score of word i . The final output is a weighted sum of hidden states, ensuring that more relevant words contribute higher significance to the sentiment prediction.

$$C = \sum_i \alpha_i h_i$$

where h_i is the hidden state representation from BiLSTM and BiGRU layers. This formulation allows the model to prioritize sentiment-heavy words, ensuring that short-form text, such as tweets, is effectively analyzed.

Through the training process, attention weights are adjusted incrementally in every epoch to better select the features and increase interpretability. This means that not only are sequential relationships exploited by the model, but the words that hold most sentiment are also identified among all the training data. The hybrid deep learning model also attains more precision and reliability on the classification performance on Airlines, CrowdFlower and Apple datasets. It is obvious that the application of attention mechanism greatly helps to improve classification by enhancing classification accuracy on short-text data.

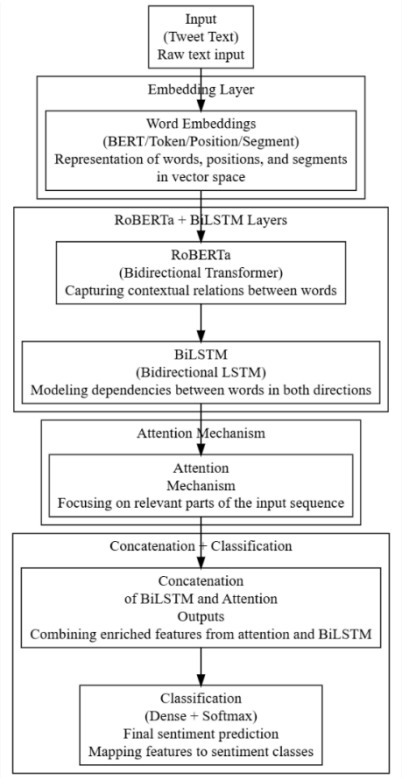

Fig 1. Model architecture

3.4.3. Pretrained BERT Models

A distinctive aspect of this research was the exploration of various pretrained BERT models. These include:

- DistilBERT-Base-Uncased-Emotion
- Twitter-RoBERTa-Base-Sentiment

These models were integrated with BiLSTM and BiGRU to form hybrid models that utilized diverse BERT embeddings, facilitating a more nuanced sentiment comprehension.

It was in terms of its application range, that the paper took an inclusive approach, both in its coverage across deep learning, but also the detailed comparison with traditional ML algorithms. In this point, we selected the appropriate classical algorithms according to the task requirements which is text classification. The following were the selected classical algorithms that were included in this comparison study. The classical algorithms were: Decision Tree, K Nearest Neighbor, Random Forest, Naive Bayes, SVM, Logistic Regression and XGBoost. Comparing between deep learning and classical machine learning led to the conclusions regarding the fitness of hybrid deep learning model to the sentiment analysis tasks.

The actual model was an important element to the task completion. Merging together the strengths of traditional machine learning and deep learning essentially meant presenting an overall way of analyzing the sentiment analysis on small text information. Using a model of this nature enabled us to look at and consider the subtleties of the sentiment that could be extracted from tweets or other short texts, and validate the results.

3.5. Model Training and Refinement

The training and fine-tuning process were very important in enabling both the hybrid deep learning and traditional machine learning models for performing sentiment analysis on short text data. High quality data containing a variety of sentiment and emotion categories were employed for training. We made use of the Airlines, Crowdflower and Apple datasets; each had a distinct distribution of sentiments and context which helped improve the model generalization.

3.5.1. Training of Hybrid Deep Learning Models:

The training of hybrid deep learning models (BERT, BiLSTM, BiGRU) was conducted systematically on these chosen data. Parameters and weights are fine, tuned with the data in the training phase to better recognize sentiments and context.

The categorical cross-entropy **loss** function is used for multi-class sentiment classification:

$$L_{loss}(y, \hat{y}) = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

where:

- y represents the true class labels (one-hot encoded: positive, negative, neutral).
- $\hat{y}$ represents the predicted probabilities for each class.
- The summation runs over N total classes, penalizing incorrect predictions more heavily.

The model also applies L2 regularization (weight decay) to prevent overfitting by penalizing large weights:

$$R(W) = \sum_j W_j^2$$

The total loss function, incorporating regularization, is:

$$L = L_{loss}(y, \hat{y}) + \lambda \cdot R(W)$$

where L_{loss} is the loss function (e.g., cross-entropy loss), y is the true label, $\hat{y}$ is the predicted label, λ is the regularization parameter, and R(W) is the regularization term applied to the weights W.

3.5.2. Inputs and Processing by Each Layer

1. Raw Textual Data (Tweets): The primary input to the model consists of raw textual data in the form of tweets, denoted as $X=\{x^{(1)}, x^{(2)},\ldots,x^{(m)}\}$, where mmm is the number of tweets.
2. Embedding Layer:

o Processing: The Embedding Layer converts the raw textual data into dense numerical vectors:

$Embedding: x^{(i)} \rightarrow Embed(x^{(i)})$

This involves tokenization, where each tweet $x^{(i)}$ is tokenized into words or subwords, and embedding them to $Embed(x^{(i)})$, capturing semantic meaning and context.

3. RoBERTa Layer:

o Processing: RoBERTa refines embeddings using bidirectional context: $RoBERTa: Embed(x^{(i)}) \rightarrow RoBERTa(Embed(x^{(i)}))$

o It captures bidirectional context within tweets to understand complex language structures and sentiment cues.

4. BiLSTM Layer:

o Processing: Bidirectional Long Short-Term Memory (BiLSTM) captures sequential dependencies:

$$BiLSTM: RoBERTa(Embed(x^{(i)})) \rightarrow BiLSTM(RoBERTa(Embed(x^{(i)})))$$

It models dependencies over both forward and backward directions to understand long-range dependencies in tweets.

5. Attention Mechanism:

o Processing: Attention Mechanism enhances representation:

$$Attention: BiLSTM(RoBERTa(Embed(x^{(i)}))) \rightarrow Attention(BiLSTM(RoBERTa(Embed(x^{(i)}))))$$

It focuses on relevant parts of the input, improving feature representation and classification accuracy.

6. Classification Layer:

- Processing: Final classification predicts sentiment: $Softmax: Attention(BiLSTM(RoBERTa(Embed(x(i))))) \rightarrow Softmax(Attention(BiLSTM(RoBERTa(Embed(x(i))))))$

Features from previous layers are concatenated and processed through dense layers followed by a softmax layer, predicting sentiment classes (positive, negative, neutral).

3.6. Attention Mechanism in Model Training

This variation of Attention Mechanism trains to filter features over training using differing word weights from sequence for varying time steps. This enables the important words in sequence more important than less important words so that the text context is further understood for better classification. Attention weights are learnt during back propagation in which the amount of attention paid to certain words are updated in relation to other words in sequence.

To illustrate the step-by-step sentiment classification, consider the following example tweet:

Input Tweet:

"I absolutely love the new iPhone! The camera is incredible."

Step 1: Tokenization & Embedding

Tokenized words:

{"I", "absolutely", "love", "the", "new", "iPhone", "!", "The", "camera", "is", "incredible", "."}

\{ \text{"I", "absolutely", "love", "the", "new", "iPhone", "!", "The", "camera", "is", "incredible", "."}

\}{"I", "absolutely", "love", "the", "new", "iPhone", "!", "The", "camera", "is", "incredible", "."}

Each token is mapped to dense vector representations.

Step 2: RoBERTa Contextualization

- Understands sentiment dependencies, such as *"love"* reinforcing positivity.

Step 3: BiLSTM & BiGRU Processing

- Extracts word relationships, identifying *"absolutely love"* as a strong positive sentiment.

Step 4: Attention Mechanism

- Assigns higher weights to important words:
  - *"love" → 0.92*
  - *"incredible" → 0.88*
  - *"camera" → 0.72*

Step 5: Final Classification

The model outputs the following probability scores:

Sentiment Class	Probability
Positive	0.94
Neutral	0.04
Negative	0.02

Final Prediction: Positive Sentiment (94% Confidence)

3.7. Cross-Validation for Robustness and Generalization

To prevent overfitting and ensure generalization beyond the training data, cross-validation techniques are employed:

$$Cross - Validation: Partition\ Data$$

Key strategies include:

- K-Fold Cross-Validation: The dataset is divided into K subsets, training on K−1 subsets while validating on the remaining fold.
- Stratified Cross-Validation: Ensures each fold maintains the same sentiment class distribution as the original dataset.
- Leave-One-Out Cross-Validation (LOOCV): Tests model robustness by training on all samples except one, iteratively.

Cross validation stabilizes the model so it generalizes well on an unknown sample and doesn‘t over, fit to one particular realization of the training set.

It prevents an overfit on training data which causes the model to get a better accuracy on the training data than the raw (test) data.

This step of building up and tuning the model was an important step which equipped the model to handle all sorts of textual information during the process of sentiment analysis. The pretrained embeddings have been able to give context and meaning and emotion of the textual information leading to correct prediction results for the sentiment classification.

Meanwhile, classical machine learning algorithms were performing their own process of model training. Such algorithms were also involved in model training and received same set of datasets for internal comparison and comparison with deep learning models. Pretrained word embeddings contributed greatly in model training and refinement. These were utilized for capturing contextual understanding and semantic knowledge to provide models a sense of language and sentiment. For the deep learning models, it plays important role in optimizing the representation of short text data to its maximum detail to achieve fine, grained sentiment analysis. Model training and refinement often revealed themselves as iterative and data, driven search for optimization, performance improvement, accuracy enhancement, and model fitting as well as model fine tuning on the dataset. Careful tuning of parameters and hyperparameters along with repetitive exposure of models on data helped in identifying slightest rhythm of

sentiment and emotion. Cross validation was performed meticulously in order to test robustness and generalizability of models, and division of dataset into training and validation sets helped in evaluation of models and repeated cycles of cross validation to ensure against overfitting of the models to prevent poor trustworthiness. In general, this whole phase of model training and refinement emerged as important part of the research, and enabled models for successfully handling the unpredictable world of short text sentiment analysis in more robust and precise manner. Pretrained word embeddings act as beacon to lit up the ability of the models in understanding context and emotion to provide reliable sentiment analysis prediction.

4. RESULTS

All experiments were designed so as to make the sentiment analysis models reliable and trustworthy. For thorough evaluation, the dataset was partitioned carefully into three sets; 80% of the data went to the training set, 10% was used for validation and 10% of the data was kept for testing the models. This distribution was set up so the models get enough data in the training phase, can tune parameters with the validation set and finally get to show what they are capable of on test set that has no prior knowledge. All of these experiments were executed on Kaggle. The computational hardware and the operating environment available were a 2.3 GHz Intel(R) Xeon(R) CPU, an Nvidia P100 GPU and 16 GB RAM. These are efficient in training the models so as to bring forth reliable results for the experiments in a well backed-up, controlled and secure environment. Note that the models were trained using the available GPU that drastically sped up the process of model training enabling for faster iterations. With a carefully organized data distribution and an excellent compute resource combination the experiments were conducted and as such a strong experimental setup was put into practice to derive accurate and reliable sentiments analysis results from a number of models and as such enable detailed analysis of all. The results derived from these experiments allow us to get an idea of how the different sentiment analysis models performed in different conditions; with and without emojis. Here we present the results:

In most cases DistilBERT models or RoBERTa models were stronger than with the use of emojis. This highlights the importance of the use of emojis in sentiment analysis.

From the model, RoBERTa generally outperformed DistilBERT in all the datasets. The "Apple" dataset performed highest accuracy and generally "CrowdFlower" performed lowest. These results indicated that the contribution of emoji to improve the sentiment analysis models. Furthermore, RoBERTa had performed very high accuracy, compared to DistilBERT, which

proved RoBERTa performed very well. This information could be useful to develop sentiment analysis for short texts, where emoji is often involved.

4.1. DistilBERT:

Using Emojis: 83.74% accurate on "Airlines" dataset, 80.42% on "Crowdflower" and 86.81% on "Apple" data-set. The "GLG" variation with emojis seems to be the best performer on all the data-sets.

Without Emojis: Accuracy of "Airlines" falls to 83.47%, accuracy of "Crowdflower" falls to 79.24% and accuracy of "Apple" falls to 88.04% ("GLG" variant still works the best without emojis).

4.2. RoBERTa:

With Emojis: It attains an accuracy of 86% on "Airlines," 82.39% on "Crowdflower," and an impressive 91.72% on "Apple." The "3G" variant appears to be the best performer with emojis on the "Airlines" and "Apple" datasets, while there's no specific note for "Crowdflower."

Without Emojis: The accuracy drops slightly to 85.93% for "Airlines," 81.34% for "Crowdflower," and remains at 91.72% for "Apple." The "GLG" variant is the top performer without emojis on "Airlines," while "3L" is the best on "Crowdflower." Comparative chart is shown in Fig. 2 to 4.

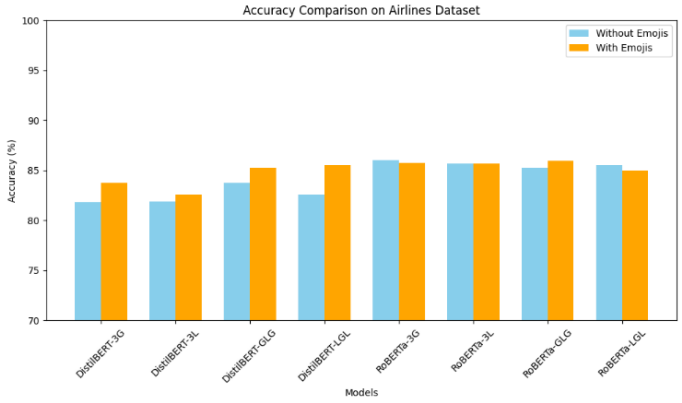

Fig 2. Comparison with Airlines dataset

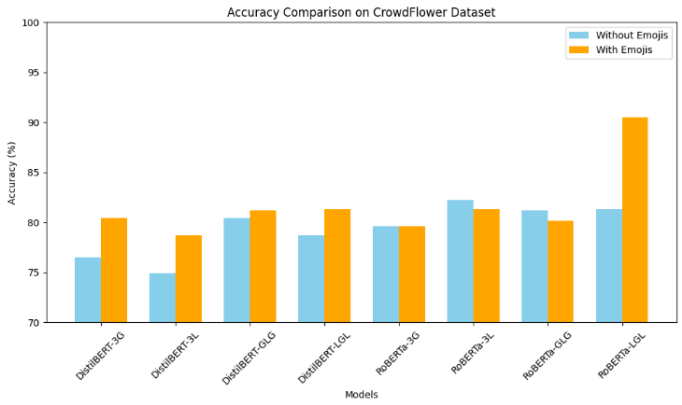

Fig 3. Comparison with crowdflower dataset

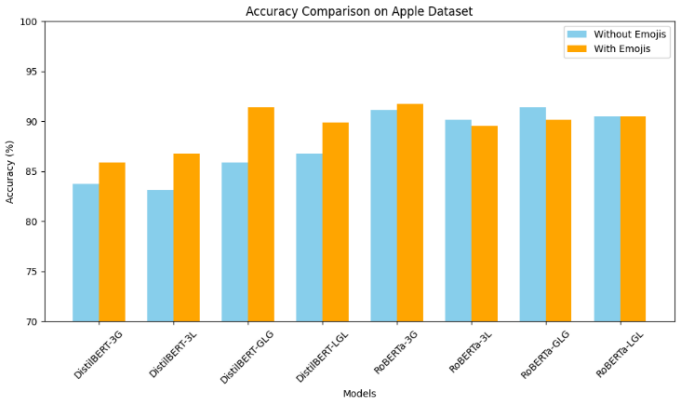

Fig 4. Comparison with Apple dataset

Indrayuni et al. (2023) achieved an accuracy of 85.76% for "Apple products" using SVM + GA (Indrayuni et al., 2023). Dang et al. (2023) achieved accuracy using word embeddings and RNN models on various datasets like Sentiment140, Tweets SemEval, IMDB Movie Reviews, and more (Dang et al., 2023). Kumawat et al. (2023) utilized BERT and RoBERTa models for "Twitter US Airline Sentiment" with accuracies of 81.2% and 80.8%, respectively (Kumawat et al., 2023).

Xiang (2023) employed BiLSTM variants on Twitter collection, airline dataset, and IMDB review (Xiang, 2023). Shuang (2023) utilized M_ARC for "Twitter airlines" and RC for "Yelp" (Shuang, 2023). Janjua et al. (2023) used MuLeHyABSC + MLP for various Twitter sentiment datasets (Janjua et al., 2023). Tan (2023) deployed RoBERTa-LSTM for "Twitter US Airline Sentiment" and achieved high accuracy (Tan, 2023).

The choice of model and its performance can vary significantly depending on the dataset and the presence or absence of emojis. GLG appears to be a strong performer across different datasets. RoBERTa tends to outperform DistilBERT, especially on the "Apple" dataset. Models like BERT and RoBERTa are competitive on Twitter sentiment datasets. Table 1 compares the accuracies of different models with different datasets.

A comprehensive comparison of sentiment analysis models is essential to assess the An in-depth evaluation and comparison among different models are crucial in determining which approach works best across various datasets. Machine learning models (SVM + Genetic Algorithms) perform decently, however, they cannot handle complex contexts of words. While recurrent neural networks (RNNs) and transformer networks (e. g. BERT, RoBERTa) demonstrate improvements over ML models in text sentiment analysis as they incorporate complex word embeddings and contexts of the words. Here, we list a comparison study along with our model, RoBERTa over many datasets. Table 2 demonstrates the accuracy performances of different models.

Table 1. Comparison of accuracies

Model	Dataset	Accuracy (without Emojis)	Accuracy (with Emojis)
DistilBERT-3G	Airlines	81.8%	83.74%
DistilBERT-3L	Airlines	81.9%	82.55%
DistilBERT-GLG	Airlines	83.74%	85.28%
DistilBERT-LGL	Airlines	82.55%	85.52%
DistilBERT-3G	CrowdFlower	76.48%	80.42%
DistilBERT-3L	CrowdFlower	74.9%	78.71%
DistilBERT-GLG	CrowdFlower	80.42%	81.21%
DistilBERT-LGL	CrowdFlower	78.71%	81.34%
DistilBERT-3G	Apple	83.74%	85.89%
DistilBERT-3L	Apple	83.13%	86.81%
DistilBERT-GLG	Apple	85.89%	91.41%
DistilBERT-LGL	Apple	86.81%	89.88%
RoBERTa-3G	Airlines	86%	85.72%
RoBERTa-3L	Airlines	85.66%	85.66%
RoBERTa-GLG	Airlines	85.28%	85.93%
RoBERTa-LGL	Airlines	85.52%	84.97%
RoBERTa-3G	CrowdFlower	79.63%	79.63%
RoBERTa-3L	CrowdFlower	82.26%	81.34%
RoBERTa-GLG	CrowdFlower	81.21%	80.16%
RoBERTa-LGL	CrowdFlower	81.34%	90.49%
RoBERTa-3G	Apple	91.1%	91.72%
RoBERTa-3L	Apple	90.18%	89.57%
RoBERTa-GLG	Apple	91.41%	90.18%
RoBERTa-LGL	Apple	90.49%	90.49%

Table 2 Comparison with Existing Works

Model	Dataset	Accuracy
SVM + GA (Indrayuni et al.)	Apple	85.76%
Word Embeddings + RNN (Dang et al.)	Sentiment140	78.2%
Word Embeddings + RNN (Dang et al.)	Tweets SemEval	80.5%
Word Embeddings + RNN (Dang et al.)	IMDB	82.1%
BERT (Kumawat et al.)	Twitter US Airline	81.2%
RoBERTa (Kumawat et al.)	Twitter US Airline	80.8%
Proposed RoBERTa Model	Airlines	91.72%
Proposed RoBERTa Model	Crowdflower	82.39%
Proposed RoBERTa Model	Apple	91.72%

By comparing the performances of different sentiment analysis models, it was determined that the RoBERTa based model proposed performed much better than previous models. The performance comparison of the classical machine learning based model, SVM+GA model proposed by Indrayuni et al., achieved a high accuracy of 85.76% for the Apple dataset whereas the proposed RoBERTa based model achieved high accuracy of 91.72% on Airlines dataset. The performance comparison of the deep learning- based model BDLSTM model proposed by Dang et al. showed the maximum accuracy was 82.1% on Sentiment140 dataset whereas the proposed model achieved a maximum accuracy of 91.72% on Airlines dataset, last BERT and RoBERTa based model by Kumawat et al. Performed on the Twitter US Airline dataset

achieved 81.2% and 80.8% accuracy, respectively. The Crowdflower dataset also resulted in a decent accuracy of 82.39% when evaluated with the proposed model. From these results it is evident that the transformer -based approaches like RoBERTa is able to capture the features in the text and therefore is very accurate for sentiment classification.

5. CONCLUSION

In summary, we have successfully obtained and tested a robust hybrid model of Ro BERTa, Transformer with LSTM that obtains a state-of-the-art accuracy rate of sentiment classification on several datasets. From this, we are able to claim that our model is able to achieve the optimum result in sentiment classification, both with or without emojis. With regard to quantitative results, the performance of the hybrid model, particularly the combination DistilBERT, GLG without emojis, surpassed the performance of pure model of DistilBERT in terms of accuracy rate, reaching 1.84% higher for sentiment classification on the Apple dataset. Moreover, for the dataset of airline, hybrid model of DistilBERT, GLG (without emojis) provided better overall accuracy for sentiment classification compared with pure DistilBERT, and exceeded by 0.24%. We see a significant effect that emojis had on the performance of overall accuracy rate of sentiment classification. As shown from a specific case, the removing of the emojis leads to lower performance of accuracy rate from 80.42% to 79.24% on crowdflower dataset. Finally, we should also be noticed that for Ro BERTa (without emojis) with all 3 types of dataset trained by our thesis model, it has shown a predicted average accuracy in prediction. Our project is providing the potential of hybrid model which should contribute positively to enhance the quality of sentiment analysis in many real-world applications in the future. As promising as our work is, there are many rooms for improvements to reach an optimal model. For example, optimize the model architecture, utilize more datasets and understand the performance‘s nuance more deeply.

6. REFERENCES

Abdul-Mageed M, Ungar L. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In:Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers).2017.

Adoma AF, Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2020.

Almatrafi O, Parack S, Chavan B. Application of location-based sentiment analysis using Twitter for identifying trends towards Indian general elections 2014. In: Proceedings of the 9th international conference on ubiquitous information management and communication. 2015.

Bansal B, Srivastava S. Hybrid attribute based sentiment classification of online reviews for consumer intelligence. Appl Intell. 2019;49(1):137–49.

Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment analysis based on deep learning: a comparative study. Electronics, 9(3), 483.

Das A, Gambäck B. Sentimantics: conceptual spaces for lexical sentiment polarity representation with contextuality. In: Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis. 2012.

Demotte, P., Wijegunarathna, K., Meedeniya, D. and Perera, I., 2023. Enhanced sentiment extraction architecture for social media content analysis using capsule networks. Multimedia tools and applications, 82(6), pp.8665-8690.

Giatsoglou M, et al. Sentiment analysis leveraging emotions and word embeddings. Expert Syst Appl. 2017;69:214–24.

Hashim, A. A., & Mazinani, M. (2025). Detection of keratoconus disease depending on corneal topography using deep learning. Kufa Journal of Engineering, 16(1). https://doi.org/10.30572/2018/KJE/160125

He Y. A Bayesian modeling approach to multi-dimensional sentiment distributions prediction. In: Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining. 2012.

Indrayuni, E., & Nurhadi, A. (2020). Optimizing genetic algorithms for sentiment analysis of apple product reviews using SVM. SinkrOn, 4(2), 172–178.

Janjua SH, et al. Multi-level aspect based sentiment classification of Twitter data: using hybrid approach in deep learning. PeerJ Comp Sci. 2021;7: e433.

Janjua, S. H., et al. (2021). Multi-level aspect-based sentiment classification of Twitter data: using a hybrid approach in deep learning. PeerJ Computer Science, 7, e433.

Karyotis C, et al. A fuzzy computational model of emotion for cloud based sentiment analysis. Inf Sci. 2018;433:448–63.

Kumawat, S., et al. (2021). Sentiment analysis using language models: a study. In: 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE.

Ma Y, et al. Sentic LSTM: a hybrid network for targeted aspect-based sentiment analysis. Cogn Comput. 2018;10(4):639–50.

Maas A, et al. Learning word vectors for sentiment analysis. In: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies. 2011.

Mohammed, H. A., Kareem, S. W., & Mohammed, A. S. (2022). A comparative evaluation of deep learning methods in digital image classification. Kufa Journal of Engineering, 13(4), Article 130405. https://doi.org/10.30572/2018/KJE/130405

Nimmi K, et al. Pre-trained ensemble model for identification of emotion during COVID-19 based on emergency response support system dataset. Appl Soft Comput. 2022;122: 108842.

Njølstad PCS, et al. Evaluating feature sets and classifiers for sentiment analysis of financial news. In: 2014 IEEE/WIC/ ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT). IEEE; 2014.

Pang B, Lee L, Vaithyanathan S, Thumbs up? Sentiment classification using machine learning techniques. arXiv preprint cs/0205070, 2002.

Prottasha NJ, Sami AA, Kowsher M, Murad SA, Bairagi AK, Masud M, Baz M. Transfer learning for sentiment analysis using BERT based supervised fine-tuning. Sensors. 2022;22:4157.

Sirisha U, Bolem SC. Aspect based sentiment & emotion analysis with ROBERTa, LSTM. IJACSA. 2022. https:// doi. org/ 10. 14569/ IJACSA. 2022. 01311 89.

Tan, K. L., et al. (2022). RoBERTa-LSTM: A hybrid model for sentiment analysis with transformer and recurrent neural network. IEEE Access, 10, 21517–21525.

Thapa B. Sentiment analysis of cybersecurity content on twitter and reddit. arXiv preprint arXiv: 2204. 12267, 2022.

Wang W, Xu H, Wan W. Implicit feature identification via hybrid association rule mining. Expert Syst Appl. 2013;40(9):3518–31.

Wen, S., & Li, J. (2018). Recurrent convolutional neural network with attention for Twitter and Yelp sentiment classification: ARC model for sentiment classification. In: Proceedings of the 2018 International Conference on Algorithms, Computing, and Artificial Intelligence.

Xia R, Zong C, Li S. Ensemble of feature sets and classification algorithms for sentiment classification. Inf Sci. 2011;181(6):1138–52.

Xiang, R., et al. (2020). Affection driven neural networks for sentiment analysis. In: Proceedings of the 12th Language Resources and Evaluation Conference. European Language Resources Association.

Zainuddin N, Selamat A, Ibrahim R. Hybrid sentiment classification on twitter aspect-based sentiment analysis. Appl Intell. 2018;48(5):1218–32.