Comparative Analysis of BERT Models for Sentiment Analysis on Twitter Data

Conference Paper · December 2023

DOI: 10.1109/ICSCC59169.2023.10335061

4 authors, including:

Nabanita Das
Techno International New Town, Kolkata, India
27 PUBLICATIONS **185** CITATIONS

SEE PROFILE

Bikash Sadhukhan
46 PUBLICATIONS **509** CITATIONS

SEE PROFILE

Comparative Analysis of BERT Models for Sentiment Analysis on Twitter Data

Arghya Sahoo
CSE, TINT
Kolkata, India
arghya.sahoo.2019@cse.tict.edu.in

Ritaban Chanda
CSE, TINT
Kolkata, India
ritaban.chanda.cse.2020@tint.edu.in

Nabanita Das
CSE, TINT
Kolkata, India
nabanita.das@tict.edu.in

Bikash Sadhukhan
CSE, TINT
Kolkata, India
ORCID: 0000-0001-5469-0711

***Abstract*— Sentiment analysis is a technique in natural language processing that entails the recognition and categorization of viewpoints articulated in written or spoken language. The utilization of Twitter as a prevalent medium for the expression of opinions and emotions has resulted in a significant focus on the analysis of sentiments from Twitter's Tweet data. Bidirectional Encoder Representations from Transformers (BERT) is a preexisting deep learning model that has demonstrated exceptional proficiency in Natural Language Processing (NLP) assignments, as in sentiment analysis. The present investigation entails a comparison of three different BERT models, namely, fine-tuned bert-base-multilingual-uncased, RoBERTuito, and RuBERT, with respect to their efficacy in performing Twitter data sentiment analysis. The models were graded according to their F1-score, recall, accuracy, and precision, utilizing a manually annotated dataset of 1,578,627 tweets. The experimental findings indicate that the RoBERTuito model exhibits superior in comparison to the other two models, attaining an accuracy score of 83.23%. The results offer valuable perspectives on the efficacy of BERT models in the context of analyzing Twitter data based on how people feel, thereby aiding scholars and professionals in selecting the most appropriate BERT model for their sentiment analysis task. The study emphasizes that the selection of the BERT model for Twitter data sentiment analysis should be contingent on the evaluation metric of interest. This study contributes to an understanding of the effectiveness of BERT models for sentiment analysis on Twitter data, which can be applied to social media monitoring and analysis.**

Keywords*— Sentiment analysis, Twitter, text data, BERT, deep learning.

I. Introduction

The advent of social media platforms has led to the emergence of Twitter as a significant channel for individuals to disseminate their viewpoints and articulate their emotions on a diverse range of subjects, encompassing politics and entertainment. The task of Twitter data sentiment analysis is a vital aspect of NLP, which entails the identification and categorization of sentiment conveyed through tweets [1]. Precisely assessing the sentiment expressed in tweets can yield significant insights into prevailing public opinion and can prove beneficial in a range of contexts, including brand management, political analysis and price prediction. Deep learning models, such as BERT, have demonstrated remarkable efficacy in a range of NLP applications, including sentiment analysis [2]. BERT models undergo pretraining on vast quantities of textual data and can be subsequently fine-tuned for particular tasks with minimal supplementary training data [3]. Therefore, they have become a popular choice for sentiment analysis on Twitter data.

However, the performance of BERT models can vary depending on the language and domain of the text data [4]. The selection of the BERT model may have an effect on the efficacy of sentiment analysis. Hence, it holds significance to conduct a comparative analysis of diverse BERT models concerning their efficacy in sentiment analysis of Twitter data. Despite the growing significance of sentiment analysis on Twitter data, there exists a paucity of research that compares the efficacy of various BERT models for this particular task. Although BERT has demonstrated remarkable proficiency in diverse NLP assignments, the selection of a BERT model may considerably influence the precision and efficacy of sentiment analysis on Twitter data. Consequently, there exists a necessity for conducting research that compares the efficacy of various BERT models in the analysis of emotions in context pertaining to Twitter data.

The current study aims to bridge this gap by conducting a comparative evaluation of three BERT models for Twitter data sentiment analysis. The motivation behind our research is to offer valuable insights into the efficacy of various BERT models in the field of opinion mining on Twitter data. Our aim is to assist professionals in selecting the most suitable model for their particular use case. This study presents an examination and comparison of the three different BERT models, namely, fine-tuned bert-base-multilingual-uncased (nlptown/bert-base-multilingual-uncased-sentiment), RoBERTuito (pysentimiento/robertuito-sentiment-analysis), and RuBERT (blanchefort/rubert-base-cased-sentiment), with respect to their emotional analysis performance. The dataset used for this analysis is the Twitter Sentiment Analysis Training Corpus, which comprises 1,578,627 tweets that have been classified as either positive (marked as 1) or negative (marked as 0). The novelty of this work lies in its comparative approach to evaluating the performance of these BERT models. This study aims to address a research gap in the existing literature and make a contribution to the advancement of sentiment analysis techniques for Twitter data. The findings of this research demonstrate that the RoBERTuito model exhibits superior performance in terms of accuracy (83.23%) compared to the other two models. The present study provides the following contributions:

1) Performs a comparative investigation of three famous BERT models utilizing empirical data.

979-8-3503-1409-0/23/$31.00 ©2023 IEEE

2) The optimal BERT model for sentiment analysis tasks is demonstrated.

3) Identify the variables that impact the efficacy of the BERT model in relation to particular tasks.

The present manuscript is constructed as follows: The subsequent section of the paper presents a thorough review of the existing body of work pertaining to sentiment analysis, Twitter data, and BERT models. The third section defines the three BERT models utilized in this research, and the fourth section describes the system model, which includes the dataset and evaluation metrics. In the subsequent section, a comprehensive examination is conducted to show the findings and analyses derived from a comparative investigation of diverse classification algorithms. Finally, the sixth section concludes the paper by discussing potential avenues for future research.

II. Literature Review

Sentiment analysis has emerged as a hot research subject in NLP, especially with the advent of social media platforms such as Twitter, which have provided a vast amount of data for analysis. Many researchers have explored the potential of machine learning (ML) and deep learning (DL) models for sentiment analysis. Recently, BERT has attracted a great deal of interest in NLP tasks, as in sentiment analysis. BERT is a pretrained deep learning model that employs attention mechanisms to identify contextual relationships between words, resulting in enhanced precision and performance.

The use of sentiment analysis on microblogging sites such as Twitter has become widespread. Bello *et al*. utilized BERT with variants such as CNN, RNN, and BiLSTM for sentiment analysis [5]. The experimental results demonstrate the effectiveness of using BERT with these variants in achieving high accuracy, recall, precision, and F1-score rates. Elankath and Ramamirtham created a Malayalam dataset of 2,000 tweets and employed a sentiment analysis BERT model based on transformers [6]. In comparison to alternative ML and DL models, the findings of the study indicated that BERT exhibited superior performance, surpassing other models by a substantial margin. The study conducted by Chintalapudi *et al*. examines the tweets posted by Indian citizens during the COVID-19 lockdown [7]. The authors employed BERT and other models to categorize the tweets into four distinct emotions: fear, sadness, anger, and joy. The BERT model yielded the highest level of accuracy at 89%, whereas the remaining three models exhibited comparatively lower levels of accuracy. Singh *et al.* discuss the growth of data on the internet, particularly on social media platforms such as Twitter, which have become popular for sharing opinions and ideas [8]. The paper focuses on sentiment analysis of coronavirus-related tweets, using web scraping to collect data and the BERT model for analysis. Results demonstrated that the BERT model performed better than conventional classification models. with an overall accuracy of 95.12%. Samir *et al.* proposed a model for sentiment analysis on Twitter using BERT technology [9]. The model fine-tunes BERT with a single layer and uses a fully connected neural network to achieve state-of-the-art results. The paper found that removing stop words and separating positive and negative tweets for topic extraction leads to better results.

However, the performance of BERT models can vary depending on the language and domain of the text data. Furthermore, it should be noted that various BERT models exhibit distinct architectures and parameters, which may influence their efficacy in the context of sentiment analysis. Hence, it is imperative to conduct a comparative analysis of diverse BERT models concerning their efficacy in Twitter data sentiment analysis.

III. Methodology

This section presents a comprehensive explanation of the operational processes behind the three BERT models utilized in the present investigation for conducting sentiment analysis on Twitter data. This study presents a comparative investigation of the three different BERT models, namely, fine-tuned bert-base-multilingual-uncased (nlptown/bert-base-multilingual-uncased-sentiment), RoBERTuito (pysentimiento/robertuito-sentiment-analysis), and RuBERT (blanchefort/rubert-base-cased-sentiment), in the context of sentiment analysis. The dataset used for this analysis is the Twitter Sentiment Analysis Training Corpus, which comprises 1,578,627 tweets that have been classified based on their sentiment. The novelty of this work lies in its comparative evaluation of the performance of these three BERT models.

A. BERT:

Several natural language processing tasks have been shown to benefit from language model pretraining. Model pretraining tasks include sentence-level tasks for predicting relationships between sentences and token-level tasks for producing output with fine-grained detail [10]. The utilization of pretrained language representations for downstream tasks has been facilitated through two established methodologies, namely feature-based and fine-tuning approaches [11]. Embeddings from Language Models (ELMo) is a feature-based NLP technique that employs pretrained representations in task-specific architectures. It generates contextualized word embeddings that determine the meaning of words based on the preceding and following words in a sentence. The development of ELMo was undertaken by researchers affiliated with the Allen Institute for Artificial Intelligence. Extensive evaluations have demonstrated its exceptional performance across many NLP tasks, such as question answering, sentiment analysis, and named entity recognition, positioning it at the forefront of the field. On the other hand, the generative pretrained transformer (GPT) employs a fine-tuning methodology wherein it incorporates just a limited number of task-specific parameters. This strategy involves training the model by fine-tuning all the pretrained parameters on subsequent tasks. GPT is a language model that creates text token by token by predicting the next token in a sequence based on the preceding ones. However, GPT is unidirectional, so it can only consider the context of preceding tokens in the sequence, which limits its effectiveness for certain NLP tasks.

We compare fine-tuning-based approaches in this study by testing with three BERT models: Transformer-based bidirectional encoder representations. The "masked language model (MLM)" pretraining target in BERT is inspired by the Cloze task and enables the model to overcome the unidirectionality constraint. The input tokens are randomly masked by the masked language model so that the original vocabulary id may be predicted. BERT is an advanced pretrained language model created by Google AI Language

[12]. It uses a bidirectional approach to language modeling and is based on the transformer architecture. which enables it to determine the context of a word within a sentence or paragraph. The BERT model has been pretrained on a large corpus of text, including the whole English Wikipedia, and has shown exceptional performance on a variety of NLP tasks, including question answering, text categorization, and named entity recognition. On some benchmark tests, it has even outperformed humans.

BERT has gained popularity in the NLP community as a tool for developing cutting-edge NLP models. Many researchers have fine-tuned BERT for specific tasks, resulting in even better results and new NLP breakthroughs. The BERT design is based on Google's transformer architecture, which was launched in 2017 for natural language processing workloads [13]. BERT extends the transformer architecture to allow bidirectional deep neural network training for language modeling. The encoder and the pretraining tasks are the two primary components of the BERT architecture.

1) *Encoder:* The encoder is made up of a stack of transformer layers, each with two sublayers. The initial sublayer consists of a multihead self-attention mechanism, which allows the model to choose and concentrate on various segments of the input sequence to generate a representation for each word in the sequence. The second sublayer is a feedforward neural network that performs a nonlinear change to the self-attention layer's output [14].

2) *Pretraining tasks:* BERT has already received training for two tasks: next sentence prediction (NSP) and Masked Language Modeling (MLM). MLM masks a portion of the input tokens at random and trains the model to distinguish between the actual tokens and the masks [15]. This enables the model to learn the connections between the various words in the phrase and their context. NSP predicts whether or whether two input sentences are sequential, assisting the model in learning the links between distinct sentences.

NLP tasks are performed during fine-tuning, task-specific inputs are fed into a pretrained BERT model, and the model's output is fed into a classifier to make the final estimate [10]. Fine-tuning allows the model to adapt to different jobs and obtain the best results possible. The input representation of BERT employs three embedding layers, as described in the following section. The final input embedding is the product of the three embeddings that preceded it. Fig. 1 depicts the block diagram for the BERT-based classification model.

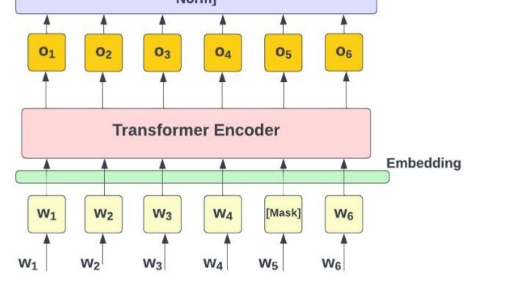
W'1
W'2
W'3
W'4
W'5
W'6
Embedding to vocab + Softmax
Classification Layer [Fully connected layer + GELU + Norm]
O1
O2
O3
O4
O5
O6
Transformer Encoder
Embedding
W1
W2
W3
W4
[Mask]
W6
W1
W2
W3
W4
W5
W6

Fig. 1 BERT model architecture block diagram

- Token Embeddings: The BERT algorithm transforms the syllables into a 768-dimensional vector representation. To serve as input representations for classification tasks and to separate input texts, [CLS] and [SEP] tokens are added to the beginning and end of the tokenized sentence. BERT utilised the WordPiece tokenizer ingeniously, which enabled it to store only 30522 words and encounter nonvocabulary words infrequently.
- Segmentation Embeddings: BERT can perform classification tasks when given two texts as input. An illustration of this would be classifying the semantic similarity of two texts. Thus, the text is concatenated and sent to BERT, which uses segment embedding to differentiate between text. The segment embedding layer consists of only two vector representations; the first (EA) is assigned to input1 tokens, and the second (EB) is assigned to input2 tokens.
- Positional Embeddings: BERT utilised learned positional embeddings by employing the transformer functions used to calculate positional encodings. This application of positional embeddings takes into account both absolute and relative positions. This is achieved by adding a sinusoidal function dependent on the position of token i in the sentence sequence and j for the position of embedding feature to the 768-dimensional vector representation of words. This results in slightly different positions for the same word in different positions. Equation (1) illustrates the functionality of BERT.

$$p_{i,j} = \begin{cases} \sin\left(\frac{i}{\frac{j}{d_{ED}}}\right) \; when\ j\ even \\ \cos\left(\frac{i}{\frac{j-1}{d_{ED}}}\right) \; when\ j\ odd \end{cases} \tag{1}$$

where ED is the embedding dimension.

B. *Model-1: Fine-tuned bert-base-multilingual-uncased [nlptown/bert-base-multilingual-uncased-statement]*

The pretrained data for this BERT model is a large corpus of text in multiple languages and is designed for sentiment analysis. It has 12 transformer layers, 768 hidden units, and 12 heads. The model uses an uncased vocabulary, meaning that all words are converted to lowercase, and there is no distinction between uppercase and lowercase letters. The input text is tokenized using WordPiece tokenization, which breaks down the text into smaller subwords. The model then includes unique tokens like [CLS] and [SEP], to append the input text to denote the start and end of a sequence. The complete sequence is represented by the [CLS] token, and its final hidden state is used for classification.

The BERT model is utilized for sentiment analysis on Twitter data through a process of fine-tuning, whereby a single linear layer is added to the model. This layer is trained to predict the sentiment of a tweet as either positive, negative, or neutral. During the process of fine-tuning, the pretrained BERT model's weights are immobilized, and solely the weights of the linear layer undergo modification. Figure 2 depicts the architecture of the fine-tuned bert-base-multilingual-uncased model.

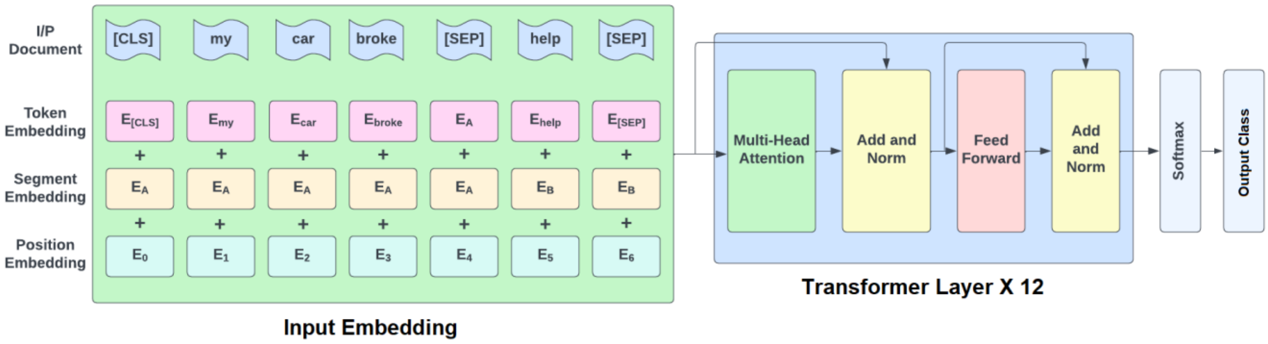

Fig. 2 Fine-tuned base-multilingual-uncased model architecture

C. *Model-2: RoBERTuito [pysentimiento/robertuito-base-uncased]*

This BERT model is a Spanish language model and is pretrained on a large corpus of Spanish text. It consists of 12 transformer layers, 768 concealed units, and 12 heads. The model uses a cased vocabulary, meaning that uppercase and lowercase letters are preserved. The input text is tokenized using SentencePiece tokenization, which is similar to WordPiece tokenization but can handle variable-length subwords. Fig. 3 illustrates the RoBERT architecture.

The pretrained model is fine-tuned for sentiment analysis on Spanish-language Twitter data by superimposing a single linear layer on top of the BERT model, which has been trained to predict whether a tweet is positive, negative, or neutral. During fine-tuning, only the weights of the linear layer are updated, while the weights of the pretrained BERT model are stationary.

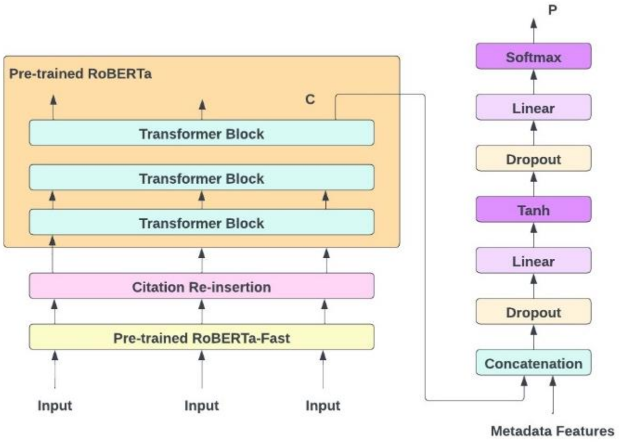

Fig. 3 RoBERTuito architecture

D. *Model-3: RuBERT [pysentimiento/robertuito-base-uncased]*

This BERT model is developed for sentiment analysis and was pretrained on a huge corpus of Russian text. It consists of 12 transformer layers, 768 concealed units, and 12 heads. The model uses a cased vocabulary, meaning that uppercase and lowercase letters are preserved. The input text is tokenized using WordPiece tokenization.

The BERT model is employed to conduct sentiment analysis on Twitter data written in the Russian language by fine-tuning an existing model. This is achieved by incorporating a singular linear layer atop the BERT model, which is specifically trained to classify tweet sentiment as either positive, neutral or negative. During the fine-tuning procedure, the parameters of the pre-existing BERT model are fixed, while only the parameters of the linear layer are adjusted.

In brief, the three BERT models operate by initially undergoing pretraining on an extensive corpus of textual data, followed by further fine-tuning on particular tasks, such as the analysis of sentiment on Twitter data. During fine-tuning, the models learn to classify the sentiment of a tweet based on its input text. The differences between the models lie in the languages and types of text data they are pretrained on, the tokenization methods they use, and the vocabulary they employ. These differences can impact the performance of the models in Twitter data sentiment analysis, which is the focus of our comparative analysis in this study.

IV. System Model:

The system model of the current work is presented in Fig. 4. In this work, Twitter data is collected from Kaggle and then preprocessed. Preprocessing is performed through the following:

- Removing stop words, punctuation, extra whitespaces, extra newlines, and duplicate entries
- Appropriate Packages should be imported for removing links, usernames and emojis//Package "re" is used to replace the links, usernames and emojis.
- Breaking down sentences into tokens
- Lemmatizing//Reduce certain words to their basic form.
- Converting all Tense to Present Tense//Only present tense is allowed.
- Converting all the words into lowercase

After preprocessing, the data are fed into three BERT models separately for sentiment analysis and accuracy calculation to evaluate the performance of each model.

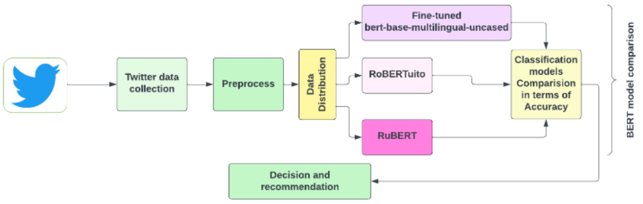
Twitter data collection
Preprocess
Data Distribution
Fine-tuned bert-base-multilingual-uncased
RoBERTuito
RuBERT
Classification models Comparision in terms of Accuracy
BERT model comparison
Decision and recommendation

Fig. 4 System model of the proposed research work

This study is carried out by comparing the aforementioned three BERT models: fine-tuned bert-base-multilingual-uncased (nlptown/bert-base-multilingual-uncased-sentiment), RoBERTuito (pysentimiento/robertuito-sentiment-analysis), and RuBERT (blanchefort/rubert-base-cased-sentiment) for sentiment analysis on the Twitter Sentiment Analysis Training Corpus dataset, which comprises 1,578,627 categorized tweets based on the TSATC dataset.

A. *Dataset collection*

The present investigation employs the TSATC dataset, which comprises 1,578,627 tweets that have been classified for Twitter Sentiment Analysis Training Corpus purposes. Each row is designated with a value of 1 to indicate a positive sentiment and a value of 0 to indicate a negative sentiment. The dataset has been compiled using information from two distinct sources.

- University of Michigan Sentiment Analysis competition on Kaggle
- Twitter Sentiment Corpus by Niek Sanders

The dataset underwent a transformation process that involved the random selection of a subset, followed by a cleaning procedure. The resulting dataset was then divided into test and training subsets, with a focus on maintaining a proportional equilibrium between positive and negative tweets within each subset. The aforementioned pair of files may be located at the following URL: https://github.com/cblancac/SentimentAnalysisBert/blob/main/data. The training subset was divided into two distinct datasets, specifically the training set, which consisted of 80% of the original data, and the validation set, which comprised the remaining 20%. The resulting dataset is composed of the previously mentioned test dataset, as well as two newly generated sub datasets. The training dataset comprises 119988 tweets, 29997 tweets for validation and 61998 tweets for testing. The BERT models used in this study are pretrained, so we use the test split in this dataset only for testing the models.

B. *Experimental Setup:*

We use Google Colab's free account to conduct our experiments. Google Colab is a cloud-based platform that provides a Jupyter notebook environment with access to a GPU, making it ideal for training deep learning models. We create a new notebook and set the runtime type to GPU.

We installed the necessary Python libraries for our experiment using pip, including transformers, pytorch, pandas, and scikit-learn. The transformer library provides access to the pretrained BERT models we will be using. PyTorch is a popular deep learning framework, and we use it to fine-tune the BERT models. Pandas is a powerful data manipulation library, and we use it to load and manipulate our datasets. Scikit-learn is a machine learning library, and we use it to evaluate the performance of our models. We enable GPU acceleration to speed up the training of our models. Google Colab provides access to a Tesla P100 GPU, which is a high-performance GPU capable of training deep learning models quickly. We ensure that we have enough RAM and disk space to run our experiments. Google Colab provides 12 GB of RAM and 100 GB of disk space by default, which is sufficient for our experiment. We write our code in Python and execute it in the Google Colab notebook.

C. *Evaluation*

Three BERT models are evaluated on a Twitter dataset for sentiment analysis utilizing classification metrics. The utilization of a confusion matrix, as depicted in the accompanying figure, is widely recognized as a reliable method for evaluating the accuracy of classification models. It quantifies the performance of such models by considering four distinct outcomes: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP refers to the accurate prediction of positive instances, TN denotes the accurate prediction of negative instances, FP signifies the incorrect prediction of positive instances, and FN represents the incorrect prediction of negative instances. Fig. 5 illustrates the confusion matrix.

Confusion Matrix		Actual	
		+VE	-VE
Prediction	+VE	TP	TN
	-VE	FP	FN

Fig. 5 Confusion Matrix

To evaluate the three specified BERT models, the accuracy, precision, recall, and F1-score evaluation metrics were calculated using a manually annotated dataset of 1,578,627 tweets. A brief description of each metric is provided below.

1) Accuracy: Measures the proportion of accurately predicted outcomes relative to the total number of predictions made.

$$Accuracy = \frac{TP+TN}{TP+TP+FP+FN} \quad (2)$$

2) Precision: The metric calculates the ratio of correctly predicted positive instances to the total number of positive instances.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

3) Recall: The metric calculates the ratio of correctly predicted positive instances to the total number of positive instances in the dataset.

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

4) F1-score: It is the harmonic mean of precision and recall that offers a balanced measure of both measurements.

$$F1 - score = \frac{2*Recall*Precision}{Recall+Precision} \quad (5)$$

V. Experimental Results and Analysis

The study evaluated three different BERT models for sentiment analysis on Twitter data, namely, fine-tuned bert-base-multilingual-uncased, RoBERTuito, and RuBERT.

RoBERTuito performed the best among the three models based on the evaluation metrics presented in the table, obtaining an accuracy of 83.23 percent, the highest precision score of 0.86, and a high recall score of 0.84. The refined bert-base-multilingual-uncased model obtained an accuracy of 81.23 percent, which is comparable to RoBERTuito but slightly lower. The RuBERT model had the lowest accuracy and precision at 78.44% and 0.75, respectively. When evaluating a sentiment analysis model, it is essential to note that accuracy is not always the most useful metric. In addition, it is crucial to consider the precision and recall scores, which indicate how accurately the model identifies positive and negative sentiments. In this instance, RoBERTuito had the highest precision and recall scores, indicating its ability to accurately categorize tweets as positive, negative, or neutral.

TABLE 1: COMPARATIVE RESULTS OF BERT MODELS

Model	Accuracy (%)	Precision	Recall	F1-Score
Fine-tuned bert-base-multilingual-uncased	81.23	0.81	0.79	0.80
RoBERTuito	83.23	0.86	0.84	0.84
RuBERT	78.44	0.75	0.75	0.78

Overall, the results indicate that RoBERTuito may be the most effective model for sentiment analysis tasks, especially when analyzing tweets in multiple languages. It is important to note, however, that the model used is determined by the task's specific requirements and the type of data being examined. These results indicate that BERT models can be used effectively for emotional analysis on social media datasets such as Twitter and that different pretrained BERT models can attain comparable performance. Nonetheless, additional experimentation with various BERT models and hyperparameters may result in an even greater performance on this endeavor.

VI. CONCLUSIONS

This research evaluated the sentiment analysis performance of three pretrained BERT models using a Twitter dataset. In particular, we trained and assessed the nlptown /bert-base-multilingual-uncased sentiment, pysentimiento/ robertuito sentiment analysis, and blanchefort/rubert-base-cased sentiment models.

Our results indicate that the nlptown/bert-base-multilingual-uncased-sentiment and pysentimiento/robertuito -sentiment-analysis models performed comparably, with respective accuracies of 81.23% and 83.23%. These results indicate that pretrained BERT models can be utilized effectually for sentiment investigation of social media datasets such as Twitter. While the efficacy of the models is comparable, there are still differences between them. The pysentimiento/robertuito-sentiment-analysis model, for instance, had slightly higher precision and recall scores than the other models, whereas the blanchefort/rubert-base-cased-sentiment model had slightly lower scores across all metrics.

This research demonstrates the viability of pretrained BERT models for sentiment analysis on social media datasets. However, there is still room for improvement, and additional experimentation with various BERT models and hyperparameters could potentially result in an even higher level of performance on this task. Future work will investigate the performance of other pretrained language models and investigate the use of domain-specific training data to enhance social media sentiment analysis. Furthermore, our research aims to explore the suitability of BERT models for various natural language processing endeavors, such as text categorization and identification of named entities.

REFERENCES

[1] L. Abualigah, N. K. Kareem, M. Omari, M. A. Elaziz, and A. H. Gandomi, “Survey on Twitter Sentiment Analysis: Architecture, Classifications, and Challenges,” in *Deep Learning Approaches for Spoken and Natural Language Processing*, V. Kadyan, A. Singh, M. Mittal, and L. Abualigah, Eds., in Signals and Communication Technology. Cham: Springer International Publishing, 2021, pp. 1–18. doi: 10.1007/978-3-030-79778-2_1.

[2] C. Colón-Ruiz and I. Segura-Bedmar, “Comparing deep learning architectures for sentiment analysis on drug reviews,” *Journal of Biomedical Informatics*, vol. 110, p. 103539, Oct. 2020, doi: 10.1016/j.jbi.2020.103539.

[3] W. de Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, and M. Nissim, “BERTje: A Dutch BERT Model.” arXiv, Dec. 19, 2019. doi: 10.48550/arXiv.1912.09582.

[4] Y. Kim *et al.*, “A pre-trained BERT for Korean medical natural language processing,” *Sci Rep*, vol. 12, no. 1, Art. no. 1, Aug. 2022, doi: 10.1038/s41598-022-17806-8.

[5] A. Bello, S.-C. Ng, and M.-F. Leung, “A BERT Framework to Sentiment Analysis of Tweets,” *Sensors*, vol. 23, no. 1, Art. no. 1, Jan. 2023, doi: 10.3390/s23010506.

[6] S. M. Elankath and S. Ramamirtham, “Sentiment analysis of Malayalam tweets using bidirectional encoder representations from transformers: a study,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 29, no. 3, Art. no. 3, Mar. 2023, doi: 10.11591/ijeecs.v29.i3.pp1817-1826.

[7] N. Chintalapudi, G. Battineni, and F. Amenta, “Sentimental Analysis of COVID-19 Tweets Using Deep Learning Models,” *Infect Dis Rep*, vol. 13, no. 2, pp. 329–339, Apr. 2021, doi: 10.3390/idr13020032.

[8] M. Singh, A. K. Jakhar, and S. Pandey, “Sentiment analysis on the impact of coronavirus in social life using the BERT model,” *Soc. Netw. Anal. Min.*, vol. 11, no. 1, p. 33, Mar. 2021, doi: 10.1007/s13278-021-00737-z.

[9] A. Samir, S. M. Elkaffas, and M. M. Madbouly, “Twitter Sentiment Analysis Using BERT,” in *2021 31st International Conference on Computer Theory and Applications (ICCTA)*, Dec. 2021, pp. 182–186. doi: 10.1109/ICCTA54562.2021.9916614.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv, May 24, 2019. doi: 10.48550/arXiv.1810.04805.

[11] Q. Wei *et al.*, “Relation Extraction from Clinical Narratives Using Pre-trained Language Models,” *AMIA Annu Symp Proc*, vol. 2019, pp. 1236–1245, Mar. 2020.

[12] “A Short Survey of Pre-trained Language Models for Conversational AI-A New Age in NLP | Proceedings of the Australasian Computer Science Week Multiconference.” https://dl.acm.org/doi/10.1145/3373017.3373028 (accessed May 06, 2023).

[13] K. Qin, G. Guo, and L. Wu, “Surface latent heat flux anomalies preceding inland earthquakes in China,” *Earthq Sci*, vol. 22, no. 5, pp. 555–562, Oct. 2009, doi: 10.1007/s11589-009-0555-7.

[14] Y. Lu *et al.*, “Understanding and Improving Transformer From a Multi-Particle Dynamic System Point of View.” arXiv, Jun. 06, 2019. doi: 10.48550/arXiv.1906.02762.

[15] S. Aroca-Ouellette and F. Rudzicz, “On Losses for Modern Language Models.” arXiv, Oct. 04, 2020. doi: 10.48550/arXiv.2010.01694.