BALKAN JOURNAL OF ELECTRICAL & COMPUTER ENGINEERING, Vol. 13, No. 4, December 2025
Research Article

A Comparative Analysis of Transformer Architectures for Sentiment and Emotion Classification

Hayrullah Temel, Erol Kına*

***Abstract*— This paper is a comparison of four Transformer model (BERT, ALBERT, T5, and XLNet) in the case of sentiment analysis of tasks based on data from social media. Used were two data sets: the X (Twitter) data set with tweets of messages related to games and the Emotion data set labeled in the anger, joy, and fear categories. Trained was the model in the same settings of preprocessing, training, as well as in the test settings for 3, 5, 7, and 10 epochs. Presented were results that indicated that the accuracy increased with the higher number of epochs. The maximum accuracies occurred in the case of the BERT model—88.63% in the case of the X data set as well as in the case of the Emotion data set, 97.05%. XLNet established great potential for long-range dependencies, and ALBERT obtained balanced performance due to lightweight architecture. On the contrary, performance of T5 was less in comparison to others. Generally, it could be inferred that Transformer architecture is superior to traditional machine learning technique in the context of sentiment analysis due to higher accuracy and better contextual understanding.**

Index Terms—**Natural Language Processing, Sentiment Classification, Transformer Models**

I. Introduction

SENTIMENT ANALYSIS, with its highly dynamic nature in the research landscape of Natural Language Processing (NLP), allows for the systematic analysis of human beings' emotional propensity and attitude in the texts that they write in online spaces. Here, it is feasible to categorize text in positive, negative, or neutral terms, and categorize more specified emotional states like happiness, anger, sadness, fear, surprise, and love, for instance. Widely employed in practical applications such as uncovering the social trends in public opinions, in analyzing customers' feedback for measuring satisfaction and complaints, in gaining insight into human beings' psychological states, and in building customized recommendation mechanisms in marketing procedures, this approach is highly significant in both scholarship and practical applications currently and in the foreseeable future [1], [2], [3], [4], [5].

Sentiment analysis trials were conducted in the beginning with very simple classification methods such as Naive Bayes and Support Vector Machines (SVM). Rising technology, however, inspired the applying of deep architecture of learning, i.e., Long Short-Term Memory (LSTM) networks as well as Convolutional Neural Networks (CNN), for such applications. Although these approaches, nonetheless, functioned fairly well, they did not capture long-distance dependencies as well as contextual subtleties, and moreover, they also required fairly deep feature engineering [6].

Today, the rapid growth of content volume in online forums has clearly revealed the inadequacies of traditional text classification methods and increased the need for more advanced, context-aware models. Here, Transformer-based models have led to a significant paradigm shift from traditional methods in emotion perception and sentiment analysis. Arrival of Transformers in fact replaced these problems, finding a strong solution and much enhancing the ability to comprehend content in textual form. Transformer-centric models were used extensively in recent years, and these in turn brought seminal break-throughs in sentiment analysis in their new age use of self-attention mechanisms as well as their bidirectioning processing capabilities [6]. Because of their context-aware representation ability [6], [7], Transformer model such as BERT (Bidirectional Encoder Representations from Transformers) and variants of it could better capture emotional nuances in text due to their representational power being enhanced through contextualization. Upon such contemplation, much another transformer models pre-trained such as RoBERTa (Robustly Optimized BERT Approach), ALBERT (A Lite BERT), DistilBERT (lighter version of BERT), and XLNet (Generalized Autoregressive Pretraining for Language Understanding) [8] were brought in based on BERT architecture.

BERT, due to its contextual learning in both directions, is an efficient basis for the work of sentiment classification tasks [9].

Hayrullah Temel is with Department of Artificial Intelligence and Robotics, Van Yüzüncü Yıl University, Institute of Science Van, Türkiye, (e-mail: hayrullah.temel07@gmail.com).
https://orcid.org/0009-0008-2340-7934
Erol Kına, is with Department of Computer Technology, Van Yüzüncü Yıl University, Özalp Vocational School, Van, Türkiye, (e-mail: erolkina@yyu.edu.tr).
https://orcid.org/0000-0002-7785-646X

*Corresponding author
Manuscript received Oct 13, 2025; accepted Nov. 10, 2025.
DOI: 10.17694/bajece.1802918

ALBERT decreases the number of parameters immensely using factorization of embeddings and cross-layer sharing of parameters methods and thus decreases the size of the model without introducing any significant degradation in performance and enhancing parameter efficiency [8], [10]. T5 (Text-to-Text Transfer Transformer) takes on all text-related tasks from the text-to-text transform point of view, and it applies to all problems as converting an input text to an output text [11], XLNet uses the permutation language modeling (PLM) strategy to efficiently calculate the bidirectional context by maximizing the expected log-likelihood across all permutations of words feasible in a sentence. Thus, it strengthens contextual understanding utilizing tokens in adjacent positions [12].

There are numerous studies in the literature demonstrating that Transformer-based models have achieved remarkable success in various text classification tasks. However, studies that directly compare different Transformer architectures under the same datasets and identical training conditions remain limited. This limitation makes it challenging to comparatively evaluate the contextual sensitivity and architectural distinctions of these models. In particular, social media–based data pose several challenges such as linguistic diversity, contextual ambiguity, and class imbalance—factors that directly influence the performance of Transformer-based models. Therefore, conducting a comparative analysis of different Transformer architectures on the same datasets is crucial not only for identifying which model performs more effectively in practice but also for providing a methodological roadmap for future research.

Here, in response to the gap in the existing literature, the current study systematically investigates the performance of four base Transformer-based models (BERT, ALBERT, T5, and XLNet) on two partially different datasets. While the first is a set of posts scraped from the X platform (also known in the past as Twitter) that contain user-generated content for gaming and that fall in three polarity categories: positive, negative, and neutral, the second is the Emotion dataset scraped from the X platform, wherein textural expression is tagged along three emotional states: anger, joy, and fear. The fact that both datasets contain class imbalance gives room for the models to be tested in more real-world settings.

Its novelty is that it applies comparative analysis of the four Transformer structure models on two social media–oriented datasets with the same preprocessing, training, and test process. From such a process, its merits and shortcomings in various emotion and sentiment levels are made apparent, and the correlation between epoch number and overfit is also assessed.

II. RELATED WORKS

Sentiment analysis or opinion mining is a primary Natural Language Processing (NLP) task whose essence is to categorize texts into sentiment classes such as positive, negative, or neutral. While dictionary-primarily based strategies, together with conventional machine studying strategies, had been extensively adopted in preliminary analysis, improvements obtained within the realm of deep studying, especially with regard to the advent of Transformer-based fashions, have elevated sentiment classification performance tremendously [13], [14]. These advancements have opened the door for Transformer architectures to become widely used in sentiment analysis, taking full advantage of the capabilities brought by deep learning. Recent progress in Transformer models has significantly reshaped how sentiment and attitude are classified in NLP. A wide range of studies have compared the performance of models like BERT, ALBERT, T5, and XLNet across various datasets and application areas, consistently showing that these advanced models outperform traditional machine learning techniques.

As an example, in one study, Hayatu et al. [15] compared some machine learning and deep learning models for sentiment detection in text, and determined that the BERT model was the most successful, with an accuracy of 88.67% and an F1-score of 0.8871. More or less decent performances were achieved by CNN and SVM models, while that of Naive Bayes was poorest. These findings indicate that deep models generally perform better in terms of accuracy than classic models, and that selecting an appropriate model needs to take into consideration criteria such as the particular requirements of the application, as well as interpretability of the model.

In another related study, Kına [3], investigated emotional responses related to mental health through the analysis of Turkish-language comments from Instagram. The study used some of the latest Transformer-based models such as XLM-RoBERTa-Large, BERTurk, and ElectraTurkish. Of these, XLM-RoBERTa-Large produced the best results with a 92% accuracy, as well as an F1 score of 90.5%. The findings indicated that Transformer-based models can effectively classify sentiment in low-resource languages such as Turkish.

Another experiment conducted in the clinical setup compared models such as BERT, RoBERTa, GPT-2, and XLNet in deciphering emotional finesse in doctor-patient conversations. Using 185 hours of clinical speech audio of Greek people, BERT scored higher than the rest in all emotion classes. RoBERTa emerged as best in the discrimination of neutral emotion, followed by GPT-2 being also best in the neutral class, while XLNet produced average scores [16].

Furthermore, Sokolová [17] also conducted a comparative study of RoBERTa, DistilBERT, mT5, byT5, and GPT models in terms of sentiment analysis of Slovaki-language datasets, such as SentiSK, Sentigrade, and a South African Slovak dataset. The models were compared in terms of accuracy, F1 score, precision, and recall for binary and multi-class sentiment classification tasks. It was revealed that GPT and mT5 were top performers in different datasets, while DistilBERT was also appreciably performing in task-oriented scenarios. As such, the study was fruitful in demonstrating that Transformer-based models can achieve more potent sentiment analysis when fine-tuned for specific languages and datasets.

In a study by Tan et al. [18], the authors assessed their suggested RoBERTa-LSTM model with IMDb, Twitter US Airline Sentiment, and Sentiment140 datasets. It was concluded that it decisively bettered all prior methods in all datasets. On the IMDb dataset, in particular, it realized an F1-score of 93% and accuracy of 92.96%, distinctly bettering the earlier GRU-based model. On the Twitter US Airline Sentiment dataset, the F1-score improved appreciably—from the earlier range of 45–72% to 91%—with data augmentation adding a 5.48% increase

in accuracy. On the Sentiment140 dataset, the model realized a 90% F1-score and attained improved accuracy by 10.6% compared to the second best, LSTM. The research proved that combining RoBERTa's strong contextual Embeddings with LSTM's ability to learn long-term dependencies constitutes a highly effective solution for sentiment analysis.

Areshey et al. [8] carried out a comparative analysis of sentiment classification of sentiment for different NLP models on Yelp review data. RoBERTa produced the best accuracy of 98.30%, followed by XLNet with 98.20%. BERT also did well with an accuracy of 97.40%, and ALBERT achieved 97.20%, with the added value of increased parameter efficiency. DistilBERT, with its compact architecture, attained impressive performance with a 96.00% accuracy, although it lagged slightly behind some of the larger models. On the whole, the results indicate that Transformer-based models—particularly BERT, RoBERTa, XLNet, and ALBERT—yield high accuracy and strong generalizability in sentiment classification applications. Researchers also pointed out the requirement for further study of the extensibility of these models to real-world data and heterogeneous application scenarios.

Branco et al. [19] conducted a detailed study using Transformer-based transfer learning methods to bridge the Portuguese-language sentiment classification gap. The study was of fine-tuning models of BERT and RoBERTa for deployment in real time on edge-devices such as Jetson Nano and Raspberry Pi, based on reviews of Portuguese restaurants. The best models were realized with an accuracy of 0.84, better than previously existing models such as PTT5 (0.82), and BERTimbau (0.8). They also realized better performances through F1-score, sensitivity, and specificity measures.

Kaur et al. [20] conducted a comparative analysis of Transformer-based architectures, including DistilBERT, RoBERTa, and XLM-RoBERTa, for sentiment analysis on social media data related to the United Kingdom's Central Bank Digital Currency (CBDC). Their findings revealed that RoBERTa, particularly when trained for three epochs, achieved the highest accuracy and demonstrated superior capability in capturing subtle linguistic nuances within financial discourse.

Almalki [21] proposed a Transformer-based multilingual sentiment and emotion detection system using the XLM-R model. The study compared XLM-R with mBERT, T5, and traditional classifiers such as SVM and Random Forest on multilingual social media data from Twitter, YouTube, Facebook, and Amazon Reviews. XLM-R achieved the highest F1-score of 90.3%, outperforming all baselines and demonstrating strong robustness in handling code-switching and cross-lingual text.

Taneja et al. [22] proposed an unsupervised Transformer-based approach using the fine-tuned DistilBERT model to perform sentiment analysis on an imbalanced women's clothing e-commerce dataset. The study addressed two subtasks—Sentiment Classification (SC) and Product Recommendation (PR)—achieving high performance with F1-scores of 0.79 and 0.85, and accuracies of 0.96 and 0.91, respectively. The results demonstrated that the proposed models outperformed traditional supervised and state-of-the-art approaches while maintaining robustness against data imbalance issues.

Ali et al. [23] pointed out sentiment analysis's promise of mining insights from reviews of e-commerce sites, experimental methodologies such as NLP, ML, and BERT. BERT was concluded to achieve accuracy of 89%, demonstrating the power of distinguishing in subtle sentiment nuances. Among them, BERT was at the top with the best of 89% accuracy. The findings also pointed out models with the capability of distinguishing subtle contextual subtleties, especially for distinguishing highly correlated sentiments such as 4 vs. 5 stars or 1 vs. 2. The research also concluded that Transformer models not only achieve outstanding suitability for practical, real-world applications but also show much promise in terms of integrating into commercial decision-support systems. On the whole, the literature clearly records that Transformer-based models overwhelmingly surpass traditional methods in accuracy as well as contextual discernment. Their repeating pattern of success through languages, datasets, and domains further reinforces their foundation role as a sentiment classification method in contemporary times.

III. METHODOLOGY

Principal goal of this research is to scientifically investigate Transformer-related model performance (BERT, ALBERT, T5, and XLNet) in text-related sentiment analysis tasks. Methodological approach includes the choice and preparation of sets, realization of the realization of preprocessing procedures, representation of mathematical foundations of used models, determination of learning procedures and tuning parameters, and specification of performance measuring metrics.

A. Datasets

There, two various sets of data (the Emotion dataset and the X (Twitter) set) were employed for measuring Transformer-based model performance for tasks in sentiment analysis based on social media. The Emotion dataset is composed of posts that users have generated and labeled on the X platform (previously Twitter). Post content is classified into three main emotion classes: anger, joy, and fear. With a total of 5.937 samples, it was chosen for use because it is an ideal multi-class emotion classification problem.

The second was also drawn from user posts from the X platform, and it was just for gaming-related posts. These posts contained user opinions and emotional states towards gaming issues and had been labeled by hand in three sentiment polarity groups: positive, negative, and neutral. This is a realistic representation of user-generated online discussion involving informal style, variability in emotional content, and ambiguity of context, and it is thus ideal for model robustness test in real-world sentiment analysis applications.

B. Preprocessing Steps

Various preprocessing methods were used to convert the raw text data in such a form that it could be processed in the model. URLs, user mentions, emoji, and special characters were first removed from the comments. All the texts were then made lowercase for uniformity, and tokenization was conducted for breaking sentences into words. Stopword removal was performed following this for the removal of words that do not

convey any meaningful information. Padding was used for converting the texts to fixed-length sequences for converting them to fixed-length sequences, and each word was changed to embedding vectors for the model to process words in numeric form. For example, the raw sentence "I love this game!" was first cleaned and tokenized into the form X = {"i", "love", "this", "game"}, where each word represents a token. This step corresponds to Eq. (1), which mathematically expresses the tokenized representation of a text instance.

Texts in the experiment had their words translated to their representations and mathematically represented just like in Eq. (1):

$$X = \{w_1, w_2, \dots, w_T\} \quad (1)$$

Here:

X: represents a single text instance. w_1: denotes the first word of the sentence, and T: refers to the total length of the sentence (i.e., the number of words).

As a result of the tokenization process, each word is represented by a d-dimensional embedding vector. For instance, in the same example, each token (e.g., "i", "love", "this", "game") is transformed into a numerical embedding such as e_1 = [0.12, –0.45, 0.63, …], forming the complete embedding representation of the sentence.

In Eq. (2), EX; denotes the embedding representation of the text instance Xin the embedding space.

Where: $E(X)$: the embedding representation of the sentence (or text instance), e_i: the d-dimensional vector of the i-th word, T: represents the total number of words in the sentence.

$$E(X) = \{e_1, e_2, \dots, e_T\}, \quad e_i \in R^d \quad (2)$$

C. Transformer-Based Architecture Overview

Transformer models aim to generate semantic and contextual text representations through different architectural designs and learning mechanisms. The four major models considered in this research—BERT, ALBERT, T5, and XLNet—are based on distinct variants of Transformer architectures. Each model incorporates unique contextual learning mechanisms, parameter optimization strategies, and representational enhancement techniques, making them suitable for various natural language processing tasks such as emotion and sentiment analysis. This section provides an overview of these Transformer-based architectures.

a) BERT (Bidirectional Encoder Representations from Transformers)

BERT is constructed from a bi-directional Transformer architecture that discovers contextual relations in both right-to-left and left-to-right directions. Its fundamental mechanism is that of self-attention, which enables the model to look at all dependencies between words in the sentence at once. The equation in Eq. (3) describes the process of computing the self-attention mechanism:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

Where:

Q: query vectors, K: key vectors, V: value vectors d_k: represents the dimension of the key vectors. Each of them is derived from the input embeddings and transformed using different weight matrices [24].

b) ALBERT (A Lite BERT)

ALBERT is a small and computationally less demanding extension of the BERT model architecture. Using parameter sharing and factorized embedding parameterization methods, the model decreases the memory footprint and attenuates the training time while retaining performance similar to that of deeper Transformer models.

c) T5 (Text-to-Text Transfer Transformer)

It performs all text-related tasks in the form of "text-to-text". It, in its encoder-decoder model, changes the input sequence in the form of an output sequence.

d). XLNet

Unlike BERT, it takes a permutation-based method of language model. Eq. (4) specifies the goal of permutation-based language model, and it is the fundamental learning rule of XLNet.

$$max_\theta \mathbb{E}_{z\sim Z_T}\left[\sum_{t=1}^{T} log\, P_\theta\left(x_{z_t}\middle|x_{z<t}\right)\right] \quad (4)\ [12]$$

Where:

max_θ:optimization of model parameters

$\mathbb{E}_{z\sim Z_T}$: the expected value over all permutations

x_{z_t}: the t-th word in the permutation

$x_{z_{<t}}$: the words preceding this word

D. Training Process and Optimization

All of the Transformer-related models in our experiment (BERT, ALBERT, T5, and XLNet) were all trained in a deep learning platform established on the basis of the PyTorch platform. Its primary goal was to experimentally compare the roles of various epoch values (3, 5, 7, and 10) in model performance and choose the best learning level subsequently.

a) Training Parameters

Directly affecting the learning efficiency and model generalization capacity is the primary hyperparameters applied in the model building process during its training phase. Presented below are key selected parameters.

Epochs: There were four varying epoch configurations used in the training procedure, namely 3, 5, 7, and 10. It helped in witnessing if there was any underfitting or if the model requires early termination for lesser epoch settings, and if it was vulnerable for overfitting for higher epoch settings. Since signs of overfitting were observed after 10 epochs, lower epoch values were preferred to achieve optimal performance.

Batch Size: 32. Less frequent use of batch sizes may result in noise in the gradient updates, causing instability in learning, and too-large batches result in high memory usage and higher computational cost, losing efficiency. Thus, 32 was selected for the batch size because it is popularly used in the literature for

its reasonable trade-off for training stability versus its use of computational efficiency.

Learning Rate: Set to 2×10^{-5}. This is within the proposed limit for Transformer-related models to achieve stable and slow convergences when doing fine-tuning.

Optimizer: AdamW, a variant of the Adam algorithm that is more efficient, was used. AdamW, incorporating a weight decay mechanism, assists in preventing overfitting and aiding in more generalizable representation of features.

Loss Function: Cross-Entropy Loss function, taken from the majority of multi-class classification tasks, was utilized. This function computes the discrepancy between the model's output class distribution and the actual distribution, and it actually leads the parameter updates. Its mathematical expression is given in Eq. (5).

$$L = -\sum_{i=1}^{N}\sum_{c=1}^{C} yi, c\ log\left(\hat{y}_{(i,c)}\right) \quad (5)$$

N: denotes the number of samples, C: represents the number of classes, $\hat{y}_{i,c}$: indicates the predicted probability of the i-th sample belonging to the c-th class [25].

b) Training Strategy

The model outputs were evaluated using an 80% training and 20% test dataset split. The training method and procedure applied in this research can therefore be described as follows:

Updates of Model Parameters: Updates of the model parameters had been carried out based on the gradients that would emanate from the backpropagation algorithm, and it was through the AdamW optimizer. It assisted in that the loss function was optimally minimized and that the learning of the model proceeded effortlessly in execution.

Loss Tracking: Loss values for training sets were noted at the end of every epoch. These notes were observed to check for the convergence behavior of the model, from which it is possible to spot the beginning of overfitting or underfitting.

Accuracy Monitoring: At the end of each epoch, accuracy values obtained from the test set were computed. These values formed the basis for the comparative performance results presented in Table I and Table II.

Generation of Confusion Matrix: Confusion matrices were produced after matching the predicted tags with the actual tags of the test dataset for each epoch. These confusion matrices made it possible to visualize the class-oriented performance of the models and to observe the effects of imbalanced class distributions on prediction accuracy.

c) Overfitting and Early Stopping

The risk of overfitting during training was systematically considered. As is widely known, at larger epoch values (for example, 10 epochs), it is possible that the training loss continues to decrease while the test loss starts to increase — a clear indication of overfitting. For this reason, the training process was organized in accordance with the following principles:

• Both training and test losses were closely monitored throughout the process.
• When an upward trend in test loss was detected, the early stopping method was applied to prevent overfitting and preserve the model's generalization capability.

d) Hardware and Computational Infrastructure

All the training procedures were in an NVIDIA GPU–accelerated setting, facilitating fast computation and less training time. It made it feasible to support the training of huge Transformer model scales and made it practical to execute numerous epoch configurations efficaciously.

E. Performance Metrics

Evaluation of the model performance was conducted with both overall and class-oriented measures to ensure that their performance is properly assessed comprehensively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Where:
TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) represent the number of corresponding classification outcomes for each category [26].

F. Confusion Matrix Generation

After the training and test process of each model, confusion matrices were computed for comparison of the predicted class outputs ($\hat{y}_i$) and ground truths (y_i). Diagonally, elements show correctly classified observations, and off-diagonal elements refer to incorrect classification. This technique was specifically used for the consideration of model performance due to class imbalance.

G. Epoch-Based Accuracy Calculation

At each epoch of the training, the accuracy value measured at its end was noted. This was measured as the number of properly classified sample in relation to the number of all the samples in the respective epoch. These epoch-oriented accuracy outcomes, in turn, were methodically tabulated and exhibited (such as in

Table I and Table II) in order to depict the performance evolution of the models at various training repetitions.

IV. EXPERIMENTAL RESULTS

A. X (Twitter) Dataset

In the sentiment analysis conducted on the X (Twitter) dataset, the performance comparisons of the Transformer-based models (BERT, ALBERT, T5, and XLNet) are presented in Table I.

TABLE I.
COMPARISON OF PRECISION, RECALL, AND F1-SCORE OF TRANSFORMER MODELS USING THE X (TWITTER) DATASET*

	Negative			Positive			Neutral				
Model	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)	Epoch	Accuracy (%)
ALBERT	82	80	81	75	80	77	71	68	69	3	76.40
ALBERT	87	84	85	84	85	85	77	78	77	5	82.76
ALBERT	82	87	84	87	80	83	77	79	78	7	82.08
ALBERT	86	87	87	90	80	85	77	86	81	10	84.38
BERT	89	91	90	87	87	87	85	83	84	3	87.31
BERT	88	91	90	90	87	89	85	86	85	5	87.98
BERT	89	92	90	89	89	89	88	84	86	7	88.63
BERT	89	91	90	93	83	88	81	90	85	10	87.70
T5	65	78	71	65	71	68	61	40	48	3	64.33
T5	66	80	72	71	66	68	62	49	55	5	66.27
T5	69	79	73	69	74	71	66	49	57	7	68.31
T5	68	80	74	70	73	71	66	49	56	10	68.30
XLNet	82	88	85	82	91	86	89	69	77	3	83.55
XLNet	88	89	88	92	87	89	83	86	84	5	87.53
XLNet	88	88	88	88	89	88	87	84	85	7	87.47
XLNet	83	91	87	91	89	90	87	80	83	10	87.07

*Each transformer model (BERT, ALBERT, T5, and XLNet) was tested at epochs 3, 5, 7, and 10.

The results showed that with an increase in the number of epochs, the overall accuracy of all the models improved in general. Among the Transformer-related models, the best performance was exhibited by BERT and XLNet. With respect to overall accuracy, the performance of the model was best when it was BERT and it was at the 7th epoch with an accuracy of 88.63%, while that of XLNet was similar and it was 87.53% at the 5th epoch. ALBERT had its best performance when it was at 84.38% accuracy at epoch 10 and it had stable performance despite its lean architecture. Comparing, the worst performance was when it was the model of T5, and its accuracy was in the range of 64% and 68%. For the negative sentiment class, the best performance was when it was the model of BERT at epoch 7, and it had 89% precision, 92% recall, and 90% F1-score. XLNet had its second-best performance when it was 88% precision, 89% recall, and 88% F1-score at epoch 5.

For the positive sentiment class, the XLNet model was prominent, having the best performance in all models at the 5th epoch with 92% precision, 87% recall, and 89% F1-score. The performance of the BERT model was balanced for this class, with 89% precision and 89% recall. The neutral sentiment class provided less strong performance for all three classes. Despite that, the highest performance for this category was for the BERT model at the 7th epoch, when it reached 88% precision, 84% recall, and 86% F1-score.

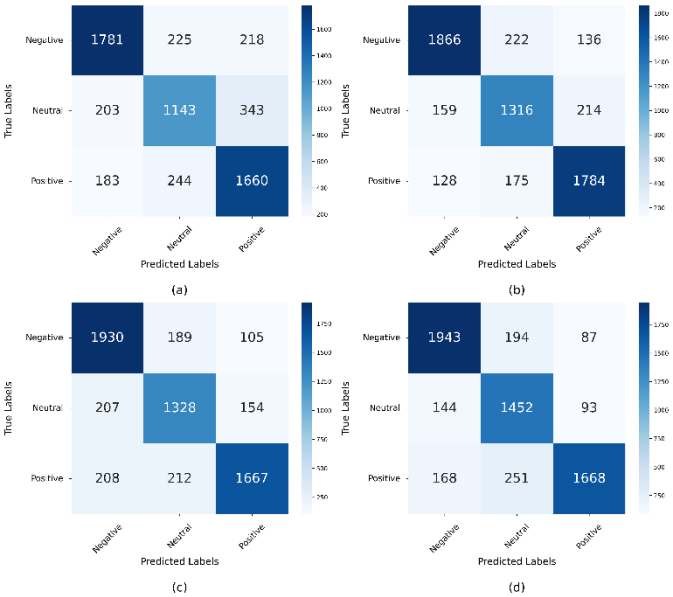

Fig.1. Confusion matrices of the ALBERT model trained on the X (Twitter) dataset for (a) 3, (b) 5, (c) 7, and (d) 10 epochs.

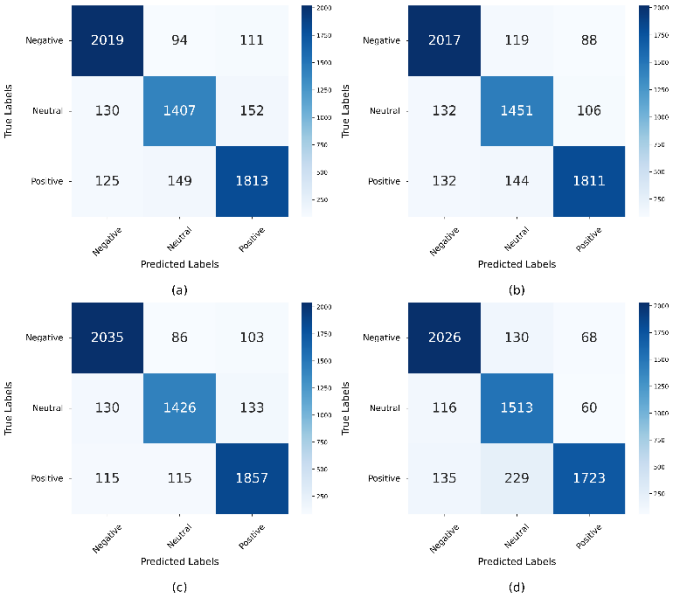

Fig.2. Confusion matrices of the BERT model trained on the X (Twitter) dataset for (a) 3, (b) 5, (c) 7, and (d) 10 epochs.

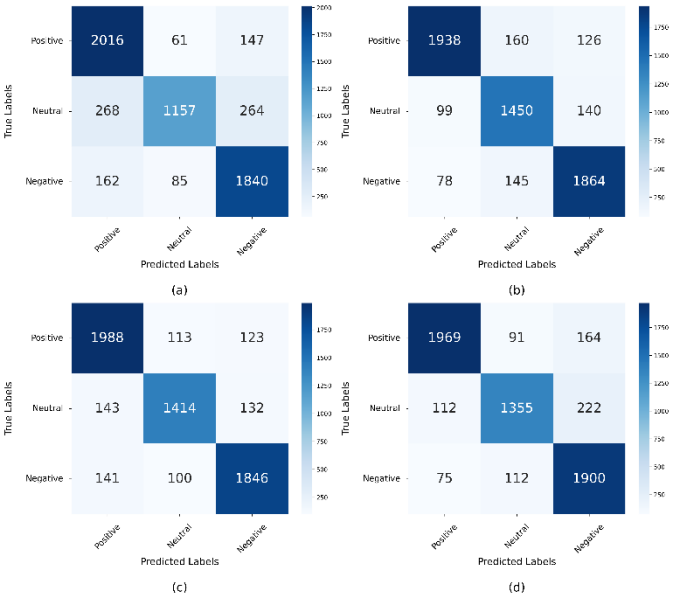

Fig.3. Confusion matrices of the XLNet model trained on the X (Twitter) dataset for (a) 3, (b) 5, (c) 7, and (d) 10 epochs.

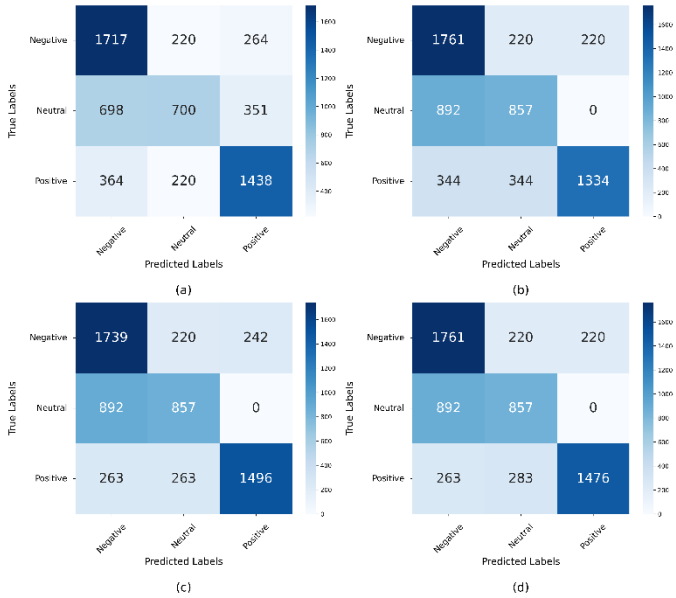

Fig.4. Confusion matrices of the T5 model trained on the X (Twitter) dataset for (a) 3, (b) 5, (c) 7, and (d) 10 epochs.

Fig. 1–4 and Table I illustrate the confusion matrices and performance metrics of the ALBERT, BERT, XLNet, and T5 models trained on the X (Twitter) dataset. All Transformer-based models achieved satisfactory results across the three sentiment categories (negative, neutral, and positive), with varying degrees of classification stability. The BERT and XLNet models demonstrated the most consistent and robust performance, effectively distinguishing between negative and positive sentiments while maintaining stable predictions for the neutral class. Minor misclassifications observed in neutral samples likely result from contextual ambiguity and overlapping expressions typical of informal social media texts. Despite its compact architecture, ALBERT performed competitively and maintained a balanced distribution of predictions across sentiment categories, confirming its effectiveness in handling context-dependent language. The T5 model, while capturing general sentiment orientation, exhibited a relatively higher confusion rate, particularly within the *neutral* class, indicating difficulty in interpreting linguistically subtle or emotionally mixed expressions.

B. Emotion Dataset

Table II presents a comparative analysis of the performance metrics (precision, recall, F1-score, and accuracy) of the Transformer-based models (ALBERT, BERT, T5, and XLNet) trained on the Emotion dataset across three emotion classes: anger, joy, and fear.

From the acquired results, accuracy tended to increase with an increase in the number of epochs in all models. Upon analyzing the entire accuracy values, the highest accuracy was obtained when the BERT model was used at epoch number 7 with 97.05%, and the least accuracy was taken when the model was the T5 and it was used at epoch number 3 with 79.88%.

In the anger class, highest performance was made by the model of BERT at the 7th iteration. At this epoch, the model had 96% precision, 96% recall, and 96% F1-score, with the best

balanced performance in this class. 96% precision, 96% recall, and 96% F1-score was made by the XLNet model at the 5th epoch, and 95% precision, 97% recall, and 96% F1-score was made by the ALBERT model at the 10th epoch.

In the joy class, the best recall value was achieved at the 10th epoch of the XLNet model. At this epoch, the model had 93% precision, 100% recall, and 97% F1-score. At the same epoch, the model of the BERT model was equally high with 99% precision, 98% recall, and 98% F1-score. Of similar remarkable performance was that of the ALBERT model, which was equally high at the same epoch with 98% precision, 99% recall, and 98% F1-score for all the epochs it was used for the classification tasks. In the fear class, the best precision value of 100% was achieved for the XLNet model, and the highest recall value of 98% was scored for both the BERT and ALBERT models. Generally, the ALBERT model had well-balanced precision, recall, and F1-score value for all the emotion classes.

TABLE II.
PERFORMANCE COMPARISON OF TRANSFORMER MODELS IN TERMS OF PRECISION, RECALL, AND F1-SCORE FOR THE ANGER, JOY, AND FEAR CLASSES IN THE EMOTION DATASET*

	Anger			Joy			Fear				
Model	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)	Epoch	Accuracy (%)
ALBERT	97	92	95	95	96	95	95	98	97	3	95.70
ALBERT	95	95	95	95	98	96	97	93	95	5	95.45
ALBERT	95	96	95	99	96	97	95	97	96	7	96.04
ALBERT	95	97	96	98	99	98	97	94	95	10	96.46
BERT	95	97	96	98	98	98	97	95	96	3	96.80
BERT	93	97	95	99	96	97	97	95	96	5	96.29
BERT	96	96	96	99	97	98	96	98	97	7	97.05
BERT	94	98	96	99	98	98	98	94	99	10	96.88
T5	76	78	77	87	82	84	77	79	78	3	79.88
T5	93	77	80	89	86	88	77	86	82	5	83.16
T5	86	84	85	93	89	91	83	89	86	7	87.28
T5	83	87	85	92	88	90	86	85	86	10	86.70
XLNet	93	97	95	99	93	96	93	95	94	3	94.78
XLNet	96	96	96	99	96	97	94	97	95	5	96.21
XLNet	92	98	95	99	97	98	98	94	96	7	96.38
XLNet	94	96	95	93	100	97	100	91	95	10	95.62

*Each transformer model (BERT, ALBERT, T5, and XLNet) was tested at epochs 3, 5, 7, and 10.

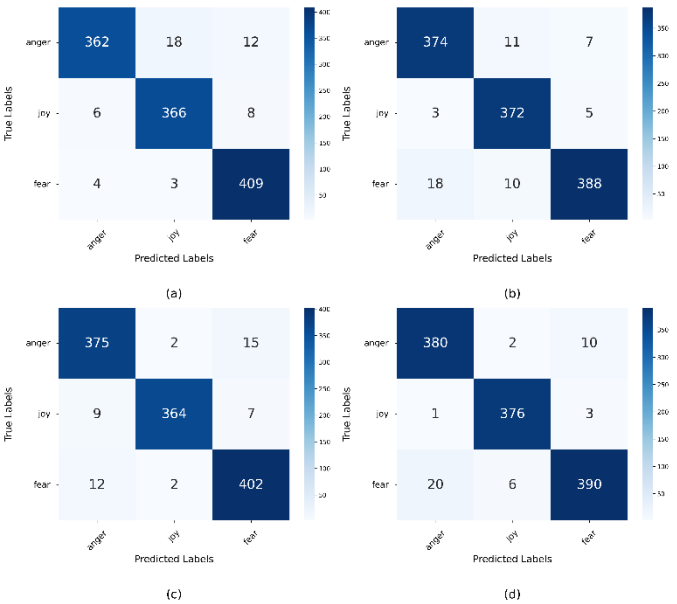

Fig.5. Confusion matrices of the ALBERT model trained on the Emotion dataset for (a) 3, (b) 5, (c) 7, and (d) 10 epochs.

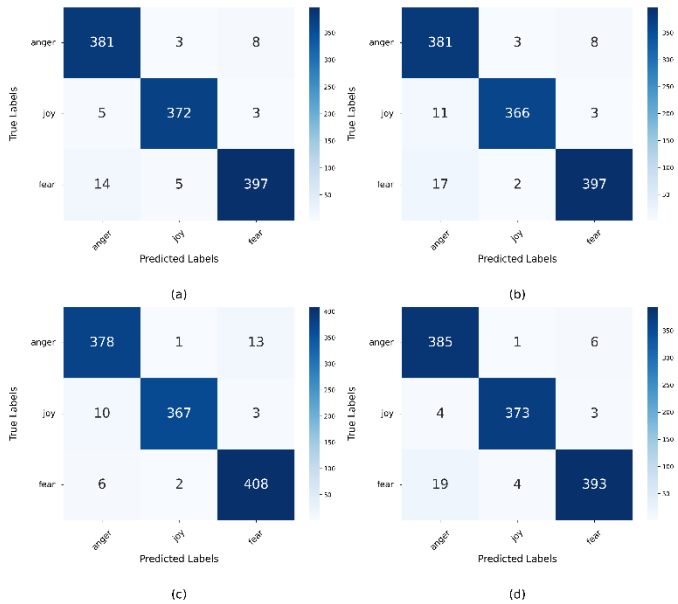

Fig.6. Confusion matrices of the BERT model trained on the Emotion dataset for (a) 3, (b) 5, (c) 7, and (d) 10 epochs.

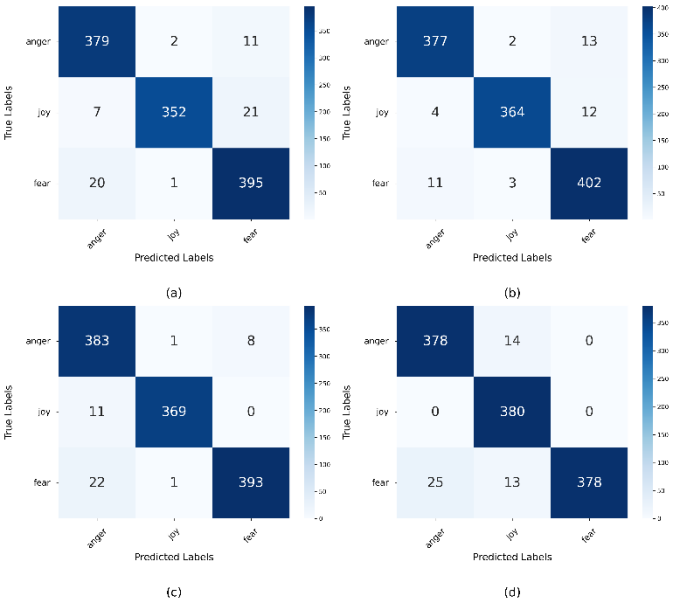

Fig.7. Confusion matrices of the XLNet model trained on the Emotion dataset for (a) 3, (b) 5, (c) 7, and (d) 10 epochs.

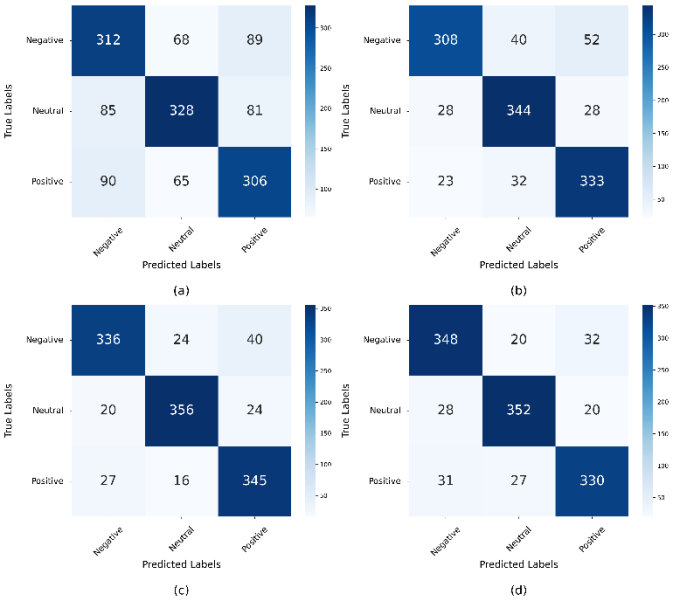

Fig.8. Confusion matrices of the T5 model trained on the Emotion dataset for (a) 3, (b) 5, (c) 7, and (d) 10 epochs.

Fig. 5–8 illustrate the confusion matrices of the ALBERT, BERT, XLNet, and T5 models trained on the Emotion dataset. All Transformer-based models achieved satisfactory performance across the three emotion categories (anger, joy, and fear). Among them, XLNet and BERT exhibited the highest accuracy, showing minimal inter-class confusion and particularly excelling in identifying the fear emotion. ALBERT, despite its lightweight architecture, demonstrated competitive performance with stable precision and recall values, effectively recognizing emotional expressions with clear contextual features. On the other hand, T5 achieved moderate but consistent results, correctly identifying most samples but showing notable overlap between anger and joy classes. This suggests that the model has some difficulty distinguishing emotions that share similar linguistic or contextual cues. Overall, the results indicate that Transformer-based architectures can effectively capture emotional nuances within text data, with XLNet and BERT outperforming the others due to their superior capacity for deep contextual representation and long-range dependency modeling.

V. DISCUSSION

The results obtained indicated that Transformer-based architectures in general possessed high values of accuracy, recall, and precision for sentiment analysis task of data from the foundation of social basis data. Out of the models compared, the BERT model performed the best in terms of accuracy across both datasets and all sentiment classes.. Additionally, it was revealed that the models were better in performance for the dataset Emotion compared to the Twitter database X. Experimental experimentation with both the X (Twitter) and Emotion sets shown that performance of the model was highly adversely affected by epochs. As the number of epochs was increased, the accuracy rates of the models tended to increase. Specifically, the best performance for all was exhibited in both sets by the model BERT, which had an accuracy of 88.63% on the X (Twitter) set and 97.05% on the Emotion set when it was in the 7th epoch.

This improvement can be attributed to BERT's bidirectional attention mechanism, which captures contextual dependencies from both directions, allowing the model to interpret subtle emotional cues and informal expressions commonly present in social media language.

These results agree with the results from the past studies in the literature, which indicated that BERT and its variants have more stable and less special results compared to the conventional machine learning techniques in capturing contextual representations efficiently. Even though accuracy is not enough for measuring the sentiment analysis models, precision and recall values, which specify the accuracy for the model in term of positive and negative emotion classification, also should not be analyzed separately.

Here, the best performance in the negative sentiment class for the X (Twitter) dataset was produced by the XLNet and BERT models. Accordingly, the BERT model achieved 89% precision and 92% recall in the 7th epoch with the best balanced and successful performance in this class. To be more precise, the XLNet model indicated a similar performance, with 88% precision and 89% recall in the 5th epoch. For the positive sentiment category, XLNet exhibited the strongest performance, achieving 92% precision and 87% recall at the 5th epoch. Compared to it, model accuracy in the neutral sentiment polarity was relatively lower than in positive and negative polarity. It can be due to the semantic vagueness of neutral terms, as they hardly contain deep emotions and thus more likely present a tougher task for the models to classify more accurately. Here, the best accuracy was achieved by the BERT model, with it achieving a precision of 88% and recall of 84% at the 7th epoch.

Given the informal, context-dependent, and semantically diverse nature of social media text, Transformer-based architectures—especially BERT and XLNet—are inherently

suitable for this sentiment analysis task, as they can effectively model contextual meaning and long-range dependencies.

Emotion dataset results also solidified the superiority of Transformer models (especially BERT, XLNet, and ALBERT) that were, in all three emotion categories, still better than all other models. BERT was particularly great at picking up contextual information, thus successfully distinguishing between both the bad emotions (e.g., anger) and the neutral emotions (e.g., joy and fear). Its best accuracy of 97.05% at the 7th iteration indicated a consistent and credible performance of the model in sentiment classification. XLNet also did exceedingly well, mostly because of its permutation-based language modeling method, through which it was well-suited for picking up long-range dependencies in text. Specifically, it attained 100% recall in the joy class and 100% precision in the fear class as it proved highly effective in picking up emotionally nuanced expressions. In conclusion, results from the Emotion dataset demonstrate that BERT, XLNet, and ALBERT all handspuriously dominate typical machine learning models when it comes to comprehending emotional context. Their proficiency in coping with complexities of sentiment analysis task further shows the merits of Transformer-based models.

Recent works by Kaur et. al [20] Almalki [21] and Taneja et al. [22] also confirm the effectiveness of Transformer-based architectures across different domains. Kaur et al. found RoBERTa to perform best in financial discourse, Almalki demonstrated the success of XLM-R in multilingual sentiment and emotion detection, and Taneja et al. showed DistilBERT's robustness in handling imbalanced e-commerce datasets. These studies further validate the generalizability of Transformer models for various text classification problems.

The present study aligns with the study findings that the BERT model is still a dependable and generalizable sentiment analysis model, particularly when used with social media data. Illustratively; in Bello et al. [27] 's study, BERT's performance was investigated when used in combination with CNN, RNN, and BiLSTM models, both with and without Word2Vec integration. The findings were that these combinations registered impressive performance in all significant metrics, such as accuracy, precision, recall, and F1-score. Likewise, Bikku et al. [28] established that BERT was better than traditional machine learning models in sentiment analysis from social media data. Their findings were that BERT was best at drawing out deep contextual subtleties and was effective in deciphering the informal, sometimes ambiguous utterances common on social sites. This virtue served BERT well in accurately predicting sentiment labels with impressive accuracy, precision, recall, and F1-scores. XLNet was the subsequent best-performing model in their comparative study.

In the experiments by Pandya [29] XLNet was revealed to immensely dominate classic algorithms including Random Forest (RF), K-Nearest Neighbors (KNN), and Support Vector Machines (SVM) in sentiment analysis of tweets. The model particularly excelled in addressing complex and ironic phrases, obtaining high F1-scores. Similarly, the XLNet-LSTM-CNN combined model of Wang and Wu [30] achieved more than 92% accuracy in tests conducted utilizing movie reviews and web comments. This paper also revealed the excellent performance of the ALBERT model, particularly on the Emotion dataset. In the paper by Ye [31] the combination of ALBERT-BiLSTM with SVM-NB overwhelmingly defeated all the models, earning higher F1-scores compared to LSTM and BERT-BiLSTM. On the dataset of ChineseNlpCorpus, the F1-score increments were 8.94%, 10.40%, and 5.10%, correspondingly. According to Ye et al., it is indicated by such improved performance that ALBERT is more successful compared to classic vector models of words in terms of retrieving textual features and providing more expressive knowledge of semantic content.

A comparative summary of the results obtained in this study and those reported in previous research is presented in Table III. The table clearly shows that Transformer-based models, particularly BERT outperform conventional machine learning methods across X datasets.

TABLE III.
COMPARATIVE SUMMARY OF THE RESULTS OBTAINED IN THIS STUDY AND THOSE REPORTED IN PREVIOUS RESEARCH

Model	Accuracy (%)	Precision (avg)	Recall (avg)	F1 (avg)	Reference
CNN	89	88	88	88	[27]*
RNN	90	90	90	90	[27]*
BiLSTM	90	90	90	90	[27]*
Word2Vec–CNN	57	57	56	56	[27]*
Word2Vec–RNN	48	48	46	45	[27]*
Word2Vec–BiLSTM	55	54	53	52	[27]*
KNN	90	88	85	87	[32]
Word2Vec–CNN-LTSM	89.8	89.38	90.93	89.17	[33]
CNN	88	87	88	87	[34]*
CNN-RNN Hybrid	82	83	82	82	[34]*
LR	75.4	75	75	75	[28]
RF	76.2	77	76	76	[28]
SVM	78.5	78	78	78	[28]
BERT-based model	86.7	87	87	87	[28]
LR	78.70	-	-	87.50	[35]
RF	79.32	-	-	87.84	[35]
RF-TD-IDF	75.89	86.76	53.65	49.84	[36]
RF-BoW	75.89	86.76	53.65	49.84	[36]
LTSM	86.68	83.06	81.06	81.23	[36]
Bert-base-uncased	88	85.77	83.33	84.01	[36]
This Study (BERT)	**88.63**	**88.6**	**88.3**	**88.3**	**Present study****

*It represents the average of the metrics for the positive, negative, and neutral sentiment classes.
**These values correspond to the epoch that yielded the highest accuracy (x dataset). It represents the average of the metrics for the positive, negative, and neutral sentiment classes

Although the proposed models achieved high performance, their accuracy was influenced by the number of epochs and dataset size, indicating potential sensitivity to overfitting. Future work will focus on optimizing training time, incorporating multilingual data, and evaluating the models on domain-specific datasets to improve their generalization ability.

Finally, it was shown through present study that Transformer-based models excelled in sentiment analysis of social media data. Of all the models that were assessed, BERT stood out with the greatest accuracy in all datasets, signifying that it is reliable and can adapt well. XLNet was differentiated by it power in

long-range dependencies, while compact but significant textual representations were pulled out by ALBERT.

These results show that Transformer models clearly outperform machine learning techniques in posting in social media. These models did not only show higher classification accuracy but also better contextual understanding, especially in short, highly emotional messages. This impressive performance suggests potential opportunities for applying such models in more advanced natural language processing applications, including irony detection, affect measuring, and multilingual sentiment analysis in the future.

ACKNOWLEDGMENT

This study was adapted from a section of the master's thesis entitled "Duygu Analizinde Transformer Tabanlı Modellerin Karşılaştırılması (Comparison of Transformer-Based Models in Sentiment Analysis)", conducted by Hayrullah Temel under the supervision of Dr. Erol Kına.

REFERENCES

[1] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Soc Netw Anal Min*, vol. 11, no. 1, p. 81, 2021.

[2] S. Ghosh, A. Ekbal, and P. Bhattacharyya, "Natural language processing and sentiment analysis: perspectives from computational intelligence," in *Computational intelligence applications for text and sentiment data analysis*, Elsevier, 2023, pp. 17–47.

[3] E. Kına, "TRANSFORMER TABANLI DUYGU SINIFLANDIRMASI İLE SOSYAL MEDYADA RUH SAĞLIĞINA İLİŞKİN TÜRKÇE YORUMLARIN ANALİZİ," *Kahramanmaraş Sütçü İmam Üniversitesi Mühendislik Bilimleri Dergisi*, vol. 28, no. 3, pp. 1499–1511, 2025.

[4] E. Kına and E. Biçek, "Machine Learning Approach for Emotion Identification and Classification in Bitcoin Sentiment Analysis," *Yüzüncü Yıl Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, vol. 29, no. 3, pp. 913–926, 2024, doi: https://doi.org/10.53433/yyufbed.1532649.

[5] E. Kına and E. Biçek, "Tweetlerin Duygu Analizi İçin Hibrit Bir Yaklaşım," *Doğu Fen Bilimleri Dergisi*, vol. 6, no. 1, pp. 57–68, 2023, doi: 10.57244/DFBD.1314901.

[6] H. Bashiri and H. Naderi, "Comprehensive review and comparative analysis of transformer models in sentiment analysis," *Knowl Inf Syst*, vol. 66, no. 12, pp. 7305–7361, Dec. 2024, doi: 10.1007/S10115-024-02214-3/FIGURES/10.

[7] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: a review of BERT-based approaches," *Artif Intell Rev*, vol. 54, no. 8, pp. 5789–5829, Dec. 2021, doi: 10.1007/S10462-021-09958-2/TABLES/18.

[8] A. Areshey and H. Mathkour, "Exploring transformer models for sentiment classification: A comparison of BERT, RoBERTa, ALBERT, DistilBERT, and XLNet," *Expert Syst*, vol. 41, no. 11, p. e13701, Nov. 2024, doi: 10.1111/EXSY.13701.

[9] S. Alaparthi and M. Mishra, "Bidirectional Encoder Representations from Transformers (BERT): A sentiment analysis odyssey," Jul. 2020, Accessed: Oct. 04, 2025. [Online]. Available: https://arxiv.org/pdf/2007.01127

[10] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[11] M. S. I. Sajol, A. S. M. J. Hasan, M. S. Islam, and M. S. Rahman, "Transforming Social Media Analysis: TweetEval Benchmarking with Advanced Transformer Models," *ISMSIT 2024 - 8th International Symposium on Multidisciplinary Studies and Innovative Technologies, Proceedings*, 2024, doi: 10.1109/ISMSIT63511.2024.10757178.

[12] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Adv Neural Inf Process Syst*, vol. 32, 2019.

[13] E. Kına and E. Biçek, "Duygu Analizinde Denetimli Makine Öğrenme Algoritmalarının Karşılaştırılmaları,(Kahramanmaraş Depremi Örneği)," *Batman Üniversitesi Yaşam Bilimleri Dergisi*, vol. 13, no. 1, pp. 21–31, 2023.

[14] E. Kına and R. Özdağ, "Deep Learning vs. Machine Learning in Sentiment Classification: A Comparative Analysis of Mobile Game Tweets from the X Platform," *Erzincan University Journal of Science and Technology*, vol. 18, no. 2, pp. 639–658, 2025.

[15] I. D. Hayatu, S. Singh, M. M. Muhammad, R. Mishra, and M. Mishra, "Emotion detection in text data: a comparative study of machine learning algorithms," *Brazilian Journal of Biometrics*, vol. 43, no. 4, pp. 1–13, Aug. 2025, doi: 10.28951/BJB.V43I4.786.

[16] M. E. Chatzimina, H. A. Papadaki, C. Pontikoglou, and M. Tsiknakis, "A Comparative Sentiment Analysis of Greek Clinical Conversations Using BERT, RoBERTa, GPT-2, and XLNet," *Bioengineering 2024, Vol. 11, Page 521*, vol. 11, no. 6, p. 521, May 2024, doi: 10.3390/BIOENGINEERING11060521.

[17] Z. Sokolová, M. Harahus, M. Sokol, E. Kupcová, and M. Pleva, "Sentiment Analysis Using Transformer Models: BERT, T5, and GPT," *Proceedings of the International Conference Radioelektronika, RADIOELEKTRONIKA*, no. 2025, 2025, doi: 10.1109/RADIOELEKTRONIKA65656.2025.11008427.

[18] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network," *IEEE Access*, vol. 10, pp. 21517–21525, 2022, doi: 10.1109/ACCESS.2022.3152828.

[19] A. Branco, D. Parada, M. Silva, F. Mendonça, S. S. Mostafa, and F. Morgado-Dias, "Sentiment Analysis in

Portuguese Restaurant Reviews: Application of Transformer Models in Edge Computing," *Electronics 2024, Vol. 13, Page 589*, vol. 13, no. 3, p. 589, Jan. 2024, doi: 10.3390/ELECTRONICS13030589.

[20] G. Kaur, · Saemundur Haraldsson, and A. Bracciali, "Comparative analysis of transformer models for sentiment classification of UK CBDC discourse on X," *Discover Analytics 2025 3:1*, vol. 3, no. 1, pp. 1–39, Jun. 2025, doi: 10.1007/S44257-025-00035-4.

[21] S. S. Almalki, "Sentiment Analysis and Emotion Detection Using Transformer Models in Multilingual Social Media Data.," *International Journal of Advanced Computer Science & Applications*, vol. 16, no. 3, 2025.

[22] K. Taneja, J. Vashishtha, and S. Ratnoo, "Transformer Based Unsupervised Learning Approach for Imbalanced Text Sentiment Analysis of E-Commerce Reviews," *Procedia Comput Sci*, vol. 235, pp. 2318–2331, Jan. 2024, doi: 10.1016/J.PROCS.2024.04.220.

[23] H. Ali, E. Hashmi, S. Yayilgan Yildirim, and S. Shaikh, "Analyzing Amazon Products Sentiment: A Comparative Study of Machine and Deep Learning, and Transformer-Based Techniques," *Electronics 2024, Vol. 13, Page 1305*, vol. 13, no. 7, p. 1305, Mar. 2024, doi: 10.3390/ELECTRONICS13071305.

[24] A. Vaswani *et al.*, "Attention is all you need," *Adv Neural Inf Process Syst*, vol. 30, no. 1, pp. 5998–6008, 2017.

[25] M. N. Razali, N. Arbaiy, P. C. Lin, and S. Ismail, "Optimizing Multiclass Classification Using Convolutional Neural Networks with Class Weights and Early Stopping for Imbalanced Datasets," *Electronics 2025, Vol. 14, Page 705*, vol. 14, no. 4, p. 705, Feb. 2025, doi: 10.3390/ELECTRONICS14040705.

[26] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf Process Manag*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/J.IPM.2009.03.002.

[27] A. Bello, S. C. Ng, and M. F. Leung, "A BERT Framework to Sentiment Analysis of Tweets," *Sensors 2023, Vol. 23, Page 506*, vol. 23, no. 1, p. 506, Jan. 2023, doi: 10.3390/S23010506.

[28] T. Bikku, J. Jarugula, L. Kongala, N. D. Tummala, and N. Vardhani Donthiboina, "Exploring the Effectiveness of BERT for Sentiment Analysis on Large-Scale Social Media Data," *2023 3rd International Conference on Intelligent Technologies, CONIT 2023*, 2023, doi: 10.1109/CONIT59222.2023.10205600.

[29] S. Pandya, "Comparative Analysis of Large Language Models and Traditional Methods for Sentiment Analysis of Tweets Dataset," *Int. J. Innov. Sci. Res. Technol*, vol. 9, no. 12, pp. 1647–1657, 2024.

[30] Y. Wang and Y. Wu, "XLNet-LSTM-CNN for text sentiment analysis," 2024.

[31] Z. Ye, T. Zuo, W. Chen, Y. Li, and Z. Lu, "Textual emotion recognition method based on ALBERT-BiLSTM model and SVM-NB classification," *Soft comput*, vol. 27, no. 8, pp. 5063–5075, Apr. 2023, doi: 10.1007/S00500-023-07924-4/FIGURES/18.

[32] F. Li, J. Li, and F. Abza, "Sentiment analysis of tweets employing convolutional neural network optimized by enhanced gorilla troops optimization algorithm," *Sci Rep*, vol. 15, no. 1, pp. 1–20, Dec. 2025, doi: 10.1038/S41598-025-85392-6;SUBJMETA.

[33] S. Tam, R. Ben Said, and Ö. Tanriöver, "A ConvBiLSTM Deep Learning Model-Based Approach for Twitter Sentiment Classification," *IEEE Access*, vol. 9, pp. 41283–41293, 2021, doi: 10.1109/ACCESS.2021.3064830.

[34] S. Riyadi, F. Daffa, C. Damarjati, and M. S. A. M. Ali, "Sentiment Analysis on Social Media Using CNN-RNN Hybrid: A Case Study of Indonesian Presidential Candidate," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, 2025.

[35] M. Liebenlito, N. Inayah, E. Choerunnisa, T. E. Sutanto, and S. Inna, "Active learning on Indonesian Twitter sentiment analysis using uncertainty sampling," *Journal of Applied Data Sciences*, vol. 5, no. 1, pp. 114–121, Jan. 2024, doi: 10.47738/JADS.V5I1.144.

[36] A. B. Alawi and F. Bozkurt, "A hybrid machine learning model for sentiment analysis and satisfaction assessment with Turkish universities using Twitter data," *Decision Analytics Journal*, vol. 11, p. 100473, Jun. 2024, doi: 10.1016/J.DAJOUR.2024.100473.

BIOGRAPHIES

Hayrullah TEMEL was born in Van in 1990. He completed his undergraduate studies in the Department of Mechanical Engineering at İnönü University in 2019 and his graduate studies in the Artificial Intelligence and Robotics Department of the Institute of Science at Van Yüzüncü Yıl University in 2025, under the supervision of Dr. Erol Kına. He develops various projects in areas such as artificial intelligence, machine learning, deep learning, and big data. He has a scientific interest in areas such as renewable energy systems, quantum computers, and chip technology.

Erol KINA was born in İstanbul. He obtained his Bachelor's degree in Computer Engineering from the Faculty of Engineering and Architecture at Çankaya University in 2005. He earned his Master's degree in Physiology from the Faculty of Medicine at Van Yüzüncü Yıl University in 2015 and completed his Ph.D. in Statistics in 2022. He fulfilled his military service in Bayburt in 2006. Between 2006 and 2008, he worked as a Systems Engineer in the private sector. In 2008, he joined Van Yüzüncü Yıl University's Özalp Vocational School as a Lecturer, where he continues to serve as an academic staff member.