Arabic ChatGPT Tweets Classification Using RoBERTa and BERT Ensemble Model

MUHAMMAD MUJAHID, Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Pakistan
KHADIJA KANWAL, Institute of CS and IT, The Women University Multan, Pakistan
FURQAN RUSTAM, School of Computer Science, University College Dublin, Ireland
WAJDI ALJEDAANI, University of North Texas, USA
IMRAN ASHRAF, Department of Information and Communication Engineering, Yeungnam University, Korea

ChatGPT OpenAI, a large-language chatbot model, has gained a lot of attention due to its popularity and impressive performance in many natural language processing tasks. ChatGPT produces superior answers to a wide range of real-world human questions and generates human-like text. The new OpenAI ChatGPT technology may have some strengths and weaknesses at this early stage. Users have reported early opinions about the ChatGPT features, and their feedback is essential to recognize and fix its shortcomings and issues. This study uses the ChatGPT tweets Arabic dataset to automatically find user opinions and sentiments about ChatGPT technology. The dataset is preprocessed and labeled using the TextBlob Arabic Python library into positive, negative, and neutral tweets. Despite extensive works for the English language, languages like Arabic are less studied regarding tweet analysis. Existing literature about Arabic tweet sentiment analysis has mainly focused on machine learning and deep learning models. We collected a total of 27,780 unstructured tweets from Twitter using the Tweepy SNscrape Python library using various hash-tags such as # Chat-GPT, #OpenAI, #Chatbot, Chat-GPT3, and so on. To enhance the model's performance and reduce computational complexity, unstructured tweets are converted into structured and normalized forms. Tweets contain missing values, URL and HTML tags, stop words, punctuation, diacritics, elongations, and numeric values that have no impact on the model performance; hence, these increase the computational cost. So, these steps are removed with the help of Python preprocessing libraries to enhance text quality and consistency. This study adopts Transformer-based models such as RoBERTa, XLNet, and DistilBERT that automatically classify the tweets. Additionally, a hybrid transformer-based model is proposed to obtain better results. The proposed hybrid model is developed by combining the hidden outputs of the RoBERTA and BERT models using a concatenation layer, then adding dense layers with "Relu" activation employed as a hidden layer to create non-linearity and a "softmax" activation function for multiclass classification. They differ from existing state-of-the-art models due to the enhanced capabilities of both models in text classification. Hybrid

M. Mujahid, K. Kanwal, and F. Rustam contributed equally to this research.

Authors' addresses: M. Mujahid, Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan, Pakistan, 64200; email: mujahidws890@gmail.com; K. Kanwal, Institute of CS and IT, The Women University Multan, Multan, 6600, Pakistan; email: khadijakanwal.6022@wum.edu.pk; F. Rustam, School of Computer Science, University College Dublin, Dublin, Ireland, D04 V1W8; email: furqan.rustam1@gmail.com; W. Aljedaani, Department of Computer Science and Engineering, University of North Texas, Texas, USA; email: wajdi.j1@gmail.com; I. Ashraf (corresponding author), Department of Information and Communication Engineering, Yeungnam University, Gyeongsan-si, Korea, 38541; email: imranashraf@ynu.ac.kr.

2375-4699/2023/08-ART204 $15.00
https://doi.org/10.1145/3605889

models combine the different models to make accurate predictions and reduce bias and enhanced the overall results, while state-of-the-art models are incapable of making accurate predictions. Experiments show that the proposed hybrid model achieves 96.02% accuracy, 100% precision on negative tweets, and 99% recall for neutral tweets. The performance of the proposed model is far better than existing state-of-the-art models.

CCS Concepts: • **Information systems → Data stream mining;**

Additional Key Words and Phrases: Arabic tweets, low-resource language, ChatGPT, OpenAI, transformer models, BERT, sentiment analysis

ACM Reference format:
Muhammad Mujahid, Khadija Kanwal, Furqan Rustam, Wajdi Aljedaani, and Imran Ashraf. 2023. Arabic ChatGPT Tweets Classification Using RoBERTa and BERT Ensemble Model. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 22, 8, Article 204 (August 2023), 23 pages.
https://doi.org/10.1145/3605889

1 INTRODUCTION

Chat generative pre-trained transformer (ChatGPT) is a deep learning-based language model, developed by Open **Artificial Intelligence (AI)**, to create human-like natural texts with amazing richness and readability [38]. ChatGPT is a variant of the most recent GPT-3 model [20] that was trained on a massive amount of data and makes predictions relevant to user input text. ChatGPT, as compared to a traditional chatbot, rejects incorrect queries and responses and recalls what the user has assumed for follow-up queries. ChatGPT provides complete details, answers, a short description, a specific response code, and written explanations on complex questions in a chat layout style. These features attract more users than traditional chatbots. ChatGPT has outperformed other online platforms in terms of popularity even after a short time of its release. ChatGPT is trained on a large amount of data; however, it is not fully valid and may sometimes provide misleading or incorrect information [47]. The ChatGPT model should be used with proper guidance to examine the results. Users are excited about the ChatGPT model, which works like a human expert. Previous GPT-3 failed to properly respond to every human query. Therefore, ChatGPT is a stronger model to answer users' questions [48].

A study conducted by Gao et al. [26] showed that the abstracts generated by the ChatGPT model are mostly identifiable both by humans and by an artificial intelligence detector. However, the ways to identify the abstract are not completely accurate. Furthermore, the authors used a small dataset for experimentation and a small number of reviewers, so the reviewers misclassified the real and ChatGPT-generated abstracts. ChatGPT is able to generate abstracts related to any journal or title. Biswas [12] demonstrates that the ChatGPT model cannot completely replace professional writers due to an insufficient level of understanding and experience in the medical field. The model's generated content may not be fully accurate or may not provide content with assessment.

Haque et al. [31] employed 10,732 Twitter ChatGPT tweets to conduct early sentiment analysis. The authors extracted important topics from tweets using a **Latent Dirichlet allocation (LDA)**-based technique and performed deep sentiment analysis on the extracted topics separately. Experiments show that early users express positive opinions and sentiments about the new artificial intelligence ChatGPT model and matching topics related to software development, entertainment, and creativity concerns. The majority of users are excited and impressed with ChatGPT's performance, but some have issues with the misuse of ChatGPT in educational fields such as online exams and assignments.

Social media platforms such as Twitter, which are mostly utilized by millions of people, create a large corpus of opinions, emotions, and attitudes. These opinions are written in the form of 140-character texts [14]. These tweets are then utilized for sentiment analysis. The users post on

a variety of topics relating to business, products, politics, and social events. Users also use social media to communicate, influence, and inform others about product details, issues, features, and so on [39]. Social media provides feedback to improve product quality and service, so businessmen can make changes according to the reviewers' opinions and feedback.

A large body of work can be found for sentiment analysis in the English language, however, sentiment analysis using the Arabic language is rather scarce. Sentiment analysis on Arabic tweets is the most challenging task for social media due to the unstructured and acoustic content of Arabic. The high number of reviews and feedback from Twitter Arabic language users require sentiment analysis to make their effective use. **Natural language processing (NLP)** is a discipline of artificial intelligence that prepares machine learning models to execute tasks related to human perspectives [17]. The primary aim of NLP is to automatically extract meaningful information from large amounts of data that contain different words with the same meanings. It is extremely difficult for humans to read, understand, and analyze a large number of tweets quickly and accurately. So, machine learning and deep learning methods can be very helpful to automatically analyze sentiments. Machine learning has been used for sentiment analysis and classification in different languages, but the low accuracy and efficacy of models are still a problem for Arabic tweets. So, to improve the accuracy and performance of sentiment analysis, this study has three main points of contribution to address these problems:

(1) A novel Transformer-based hybrid model is proposed for Arabic tweet sentiment analysis that employs **bidirectional encoder representations from Transformer (BERT)** and **robustly optimized BERT pretraining approach (RoBERTa)**. The ChatGPT OpenAI-related tweets dataset is scraped from Twitter. Different preprocessing steps are applied to unstructured tweets and annotation is carried out into positive, negative, and neutral sentiments.
(2) Two feature engineering techniques, including a **bag of words (BoW)** and **term frequency-inverse document frequency (TFIDF)**, are used for the extraction of key features from the tweets.
(3) Many machine learning and deep learning models are employed to analyze the robustness of the proposed model, including **K nearest neighbor (KNN), logistic regression, (LR), support vector machine (SVM), convolutional neural network (CNN), long short-term memory (LSTM), Bidirectional LSTM (BiLSTM)**, and **gated recurrent unit (GRU)**.

This article is further organized into four sections. The literature review is described in Section 2, while the proposed model is discussed in Section 3. Section 4 discusses experimental results and analysis. Section 5 concludes this study.

2 LITERATURE REVIEW

Sentiment analysis for social media content has become very important during recent years for opinion mining in various fields and applications [18]. Sentiment analysis is very famous for numerous platforms on the web, including forums, blogs, and e-commerce as well as social networks namely, Facebook, Twitter, YouTube, LinkedIn, and many more. However, predominantly existing works focus on sentiment analysis of English tweets.

Arabic sentiment analysis is the most challenging sentiment analysis method due to the rich morphology of the Arabic language and casual noisy content. The Arabic opinion analysis methods are attaining more significance due to the increased ratio of comments and feedback from Arabic users on several platforms [5]. Many researchers have contributed to developing methods for sentiment analysis and defined the process to detect the sentiment of various languages. Moreover, three types of Arabic are described, such as classical Arabic, which is the language of the

Quran (Islamic Holy Book), **dialectical Arabic (DA)**, and **modern standard Arabic (MSA)** [30]. There are three sentiment analysis stages named as the aspect stage, sentence stage, and document stage. In addition, the usage of these stages is an extremely challenging research field that includes various complicated tasks. There are a few more interesting topics in this research field, namely, lexicon creation, subjectivity classification, aspect-level sentiment classification, and opinion spam detection [41].

Reference [32] performs sentiment analysis on Arabic tweets using deep learning and a troupe system. Keeping in view the complexity of the Arabic language, the authors investigate various deep learning models to increase the accuracy of Arabic sentiment classification. The **Arabic Sentiment Tweets dataset (ASTD)** is used for experiments indicating a 64.46% F1 score with a fusion of CNN and LSTM. In Reference [13], SentiGAN is used to enhance the low-size dataset by generating a diversity of 12 various DA attained from the MADAR database. Experimental results demonstrate improved classification results when applying generated dataset.

Similarly, Reference [9] uses a pre-trained GPT2 approach for Arabic corpus. The experimental results report a 98% accuracy to detect the model-generated synthetic text. In Reference [4], more investigation is carried out to attain feelings from the Arabic tweets using an attention mechanism. In addition, the researchers proposed a cross-breed arrangement that combines semantic direction with an artificial intelligence technique to depict the extremity of Arabic tweets. For this, a lexical model is used to characterize the tweets in a solo manner, and the static virtual machine model secures the lexical classifier results. Experimental results show an accuracy of 84.01%. Reference [6] performs sentiment analysis of 25,000-plus Arabic tweets. To perform sentiment analysis, approximately 350,000 tweets are collected, of which 25,000-plus are annotated using crowdsourcing. The majority voting technique is applied for classification, which shows promising results.

Several works have been contributed to the Arabic language, including Arabizi [19], managing negations [8], and Arabic dialects [3, 23]. The experiments are performed using three built-in approaches in Rapid miner. The researchers attained encouraging results using this technique. Similarly, Reference [51] presented a **tweets sentiment analysis model (TSAM)**, which is based on the lexicon technique. In Reference [2], a technique to detect hate emotions in Arabic tweets is presented. Five classes of sentiments are used for experiments, including "none," "racism," "religious," "general," and "sexism hate." Moreover, the authors performed a comparison with four deep learning approaches, including CNN+LTSM, LTSM, CNN+GRU, and GRU. Experiments are performed using a dataset of 11,000 tweets. Results indicate improved results using CNN+LTSM, which shows a 72% accuracy.

Along the same directions, Reference [1] used a deep learning approach to evaluate reviews from an Arabic book reviews dataset known as **large-scale Arabic book reviews (LABR)**. The dataset consists of 16,448 reviews and contains positive and negative classes. The authors employed LSTM and its different variants using different output and batch sizes. Experimental results show an accuracy of 82% when the LSTM is used with 50 outputs and a 256 batch size. Similarly, the authors applied three deep learning models in Reference [15] to perform sentiment analysis on a 40,000 Arabic tweets dataset. Experimental results report an 88.05% accuracy using LSTM. In Reference [7], the authors used the **discriminant polynomial Naive Bayes (DMNB)** technique using 4-gram stemming, tokenizer, inverse document frequency, and word frequency for sentiment analysis. Arabic dataset containing 2,000 tweets is used for experiments showing an accuracy of 87.5% with DMNB classifier.

Analysis of the above-discussed research works explores several aspects. First, the number of research works on Arabic tweet analysis is rather few compared to other languages, especially English. Second, machine learning and deep learning models are utilized for the most part, and the transformers model is not studied very well. Third, the reported accuracy is rather low for Arabic

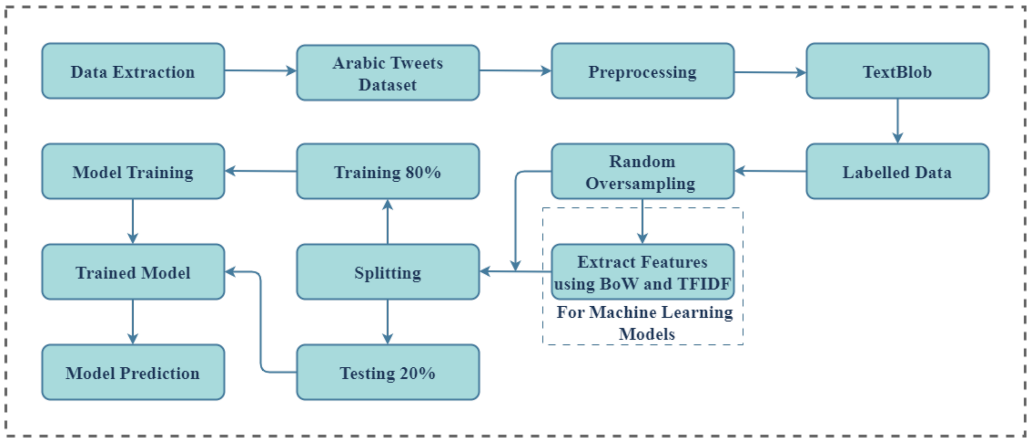

Fig. 1. Workflow of adopted methodology for Arabic tweets sentiment analysis.

Table 1. Sample of Arabic Tweets Used in This Study

	Arabic Tweets
1	تبرمجه على طريقة الرد على العملاء و الاجابه على اسئلتهم بطريقه بسيطه و سهله
2	تأتي اختبارات المنتج بعد اجتماع شامل عقد مؤخرا حيث أثار الموظفون مخاوف بشأن الميزة التنافسية للشركة في الذكاء الاصطناعي
3	شرح إنشاء الرد التلقائي في الووردبريس لدعم تجربة المستخدمين والزوّار https://t.co/aThvLR0lnp,0,0
4	تحديداً ٨ مساء بتوقيت مصر ام الدنيا https://t.co/HYiewV645G,0,0

text sentiment analysis. This study considers these aspects and investigates the use of BERT and RoBERTa in this regard.

3 MATERIALS AND METHODS

This study presents a sentiment classification approach for Arabic tweets. Figure 1 shows the workflow diagram of the adopted methodology. Starting with data collection, it follows the preprocessing, tweets annotation, and oversampling. Later, features are extracted using BoW and TFIDF to train the models that are tested using the unseen data.

3.1 Arabic Tweets Extraction and Preprocessing

ChatGPT Arabic tweets dataset is extracted from Twitter using key words such as #ChatGPT, OpenAI, Chatbot, and so on, in the Arabic language. Table 1 shows the sample Arabic tweets collected from Twitter.

The collected tweets are unstructured and contain unnecessary and redundant information. Text data preprocessing is very important for NLP tasks and is commonly utilized to enhance the capability and decrease the computational cost of machine learning models. It is used to convert textual data into a structured format using a variety of tasks. The structured and consistent data is then utilized by the models for training. Preprocessing has a great impact on the performance of NLP-related tasks [46]. So, in this study, preprocessing is used to remove HTML tags, URLs, punctuation, stop words, and so on, from the tweets. The HTML tags and URLs increase the text length and do not contribute to the model's training. Similarly, numerical values in the tweets are removed to reduce computational complexity. A few samples of preprocessed Arabic tweets are presented in Table 2.

Table 2. Sample of Preprocessed Arabic Tweets

	Preprocessed Arabic Tweets
1	تبرمجه طريقة الرد العملاء الاجابه اسئلتهم بطريقه بسيطه سهله
2	تأتي اختبارات المنتج اجتماع شامل عقد مؤخرا أثار الموظفون مخاوف بشأن الميزة التنافسية للشركة الذكاء الاصطناعي
3	شرح إنشاء الرد التلقائي الووردبريس لدعم تجربة المستخدمين والزوار
4	تحديدا بتوقيت مصر ام الدنيا

Table 3. Preprocessed Sample Tweets Are Taken from the Dataset for BoW Features

Sentence	يقدر	طريقة	تحديدا	الدنيا	بتوقيت	مصر	ام
Sentence-1	1	1	1	1	0	0	0
Sentence-2	0	0	1	1	1	1	1

3.2 Sentiment Analysis

After text preprocessing, sentiment analysis is performed using the TextBlob approach. TextBlob is a Python **natural language toolkit (NLTK)** library used to label text data into positive, negative, and neutral sentiments. It is widely used for speech tagging, language translation, and sentiment classification. NLTK provides easy access to millions of lexicon resources, and TextBlob can be used to perform complex operations on text data [29]. TextBlob returns the polarity and subjectivity of a given text. Polarity is between [−1 and 1], where −1 indicates the negative sentiment and 1 indicates the positive sentiment. Subjectivity refers to 0 and 1 and it identifies the feelings, emotions, and subjective information rather than the objective perspectives.

3.3 Feature Engineering

Two well-known feature engineering approaches are used in Arabic sentiment analysis to convert text information into numeric-vector representation. The BoW is used in natural language processing and information retrieval technique. The BoW model ignores the grammar and structure of words and only considers the frequency of words in the document. Training machine learning models need fixed-length input, so BoW can convert variable length input into fixed length input for models [40]. Table 3 shows the preprocessed tweets for BoW features.

- Sentence-1= يقدر طريقة تحديدا الدنيا
- Sentence-2= تحديدا بتوقيت مصر ام الدنيا

TFIDF is another well-known and widely utilized feature engineering approach for NLP tasks [22]. TFIDF is a slight modification of the BoW model and overcomes the limitations faced by the BoW approach. This approach does not only consider the frequency of words but also the high and low-level words and their meanings. TFIDF may be calculated with the following equation:

$$TF(t,d) = \frac{n_{T_D}}{N_{(T,D)}}, \tag{1}$$

where n_{T_D} indicate the number of occurrences of term T in a document D and $N_{(T,D)}$ represents the total terms in document D. **Inverse document-frequency (IDF)** can be calculated as

$$IDF = log\frac{T,D}{n_D}, \tag{2}$$

Table 4. Hyperparameters Tuning of Machine Learning Models

Model	Hyperparameters tuning
LR	random_state = 50, solver = "saga", multi_class = "multinomial", penalty = "l1"
DT	random_state = 150, max_depth = 150
RF	n_estimators = 50, random_state = 20, max_depth = 50
KNN	n_neighbors = 5
SVM	kernel = "sigmoid", C = 4.0, random_state = 50

where T, D shows total documents in the corpus, while n_D indicates the number-of-documents in T term. TFIDF can be obtained by multiplying TF and IDF:

$$TFIDF = (TF).(IDF). \tag{3}$$

Two important feature engineering techniques (BoW and TFIDF) are utilized in ML models for certain NLP tasks. These techniques are not applied to the transformer-based Ensemble RoBERTa and BERT models. Because transformers serve a dual purpose by functioning both as feature extractors and for classification. There is no need for manual feature engineering for transformer models. However, in ML, we need extensive feature engineering techniques to extract relevant and important features from the tweets. The RoBERTa and BERT ensemble models used Pretrained context-based word embeddings with attention mechanisms to sustain semantic information and used tokenized special text as an input for the BERT model in Arabic text analysis. The BoW and TFIDF do not support these representations for semantic information. Therefore, we cannot apply BoW and TFIDF techniques to BERT transformers, because it would lead to the loss of semantic information. Following are the limitations faced by the BoW approach and TFIDF addresses these limitations:

(1) The BoW technique considers only the frequency of a word, and words such as "the," "is," or "and" have different frequencies in the vocabulary that do not capture semantic words. TFIDF considers both frequency and inverse document frequency at the same time to capture important information.
(2) BoW gives equal importance to all words, but TFIDF gives less importance to common words and full importance to informative words.
(3) BoW treats both lengthy and short sentences equally. But TFIDF gives importance to each term separately by balancing the document.
(4) In BoW, a term's frequency determines its significance, assuming that words with a higher frequency are more significant. This assumption might not always be accurate. However. this issue is resolved by TF-IDF, which gives terms a deeper representation of their importance by scaling them in accordance with their frequency in the document (TF) and rarity in the corpus (IDF).

3.4 Fine-tuned Machine and Deep Learning Models

This study also uses different machine learning models for sentiment analysis. For this purpose, SVM, RF, LR, DT, KNN, RNN, CNN, GRU, BERT, RoBERTa, DistilBERT, XLNet, and LSTM are used. These models are fine-tuned to optimize performance. The hyperparameters tuning of machine learning models is presented in Table 4.

Support Vector Machine. SVM is employed for classification, regression, and different other tasks in various research fields. SVM divides the sample data into various classes with a set

of hyperplanes in c-dimension space, where c is employed for features [49, 50]. It performs classification to pick the "best fit" hyperplane that is employed to differentiate among classes. The model applies a "sigmoid" kernel, which is often used when the database has various features. The highest speed and better performance with a limited number of samples are the advantages of SVM. This study uses an SVM classifier with three hyperparameters, namely, "sigmoid" kernel, C regularization, and a random_state of 50.

Random Forest. RF classifier is a tree-based model and is used to produce specific predictions by combining many weak learners [45]. In RF, the bagging method is applied where several decision trees are employed during the training with different bootstrap samples [10]. These bootstrap samples are derived with sub-sampling of the training dataset using replacement. RF is also known as the attribute selection model. In ensemble classification, several models are trained and the results are pooled using a voting process. RF can be defined as

$$S = modeN1(y), N2(y), \ldots, Nt(y), \tag{4}$$

$$S = mode \sum_{r=1}^{R} (Nt(y)), \tag{5}$$

where S is used for final prediction with majority decision trees and $modeN1(y), N2(y), \ldots, Nt(y)$ is used for decision trees that take part in the process of production. This study applies RF with three hyperparameters. The t_estimtrs parameter is used with a value of 50 indicating that RF generates 20 decision trees. Additionally, the d mx_dpth hyperparameter is applied using 50, which is used to limit the decision tree to grow to a maximum of 50 levels to reduce the over-fitting and complexity.

Logistic Regression. LR is a supervised machine learning model, often used for classification problems [27]. For this, target variables are fixed and LR is used as the first choice of classification. LR is applied to manage the relationship between independent variables and dependent variables categorically by estimating probability using a logistic function. The logistic function is a normally sigmoid curve and can be defined as

$$Y = \frac{u}{1 + T^{(-k(v - v_0))}}, \tag{6}$$

where T is used for Euler number, v_0 is the value of the sigmoid mid-point, u is the maximal value for curve, and k is applied for curve steepness.

LR shows better performance for binary classification and determines the best performance to classify text [27]. This study uses LR with four hyperparameters to obtain the best results. Additionally, the "saga" techniques are employed to optimize the results.

Decision Tree. DT has a tree-like structure and is a commonly used model for classification [36]. Using this approach, connection points among the branches signify the conditions for discriminating, and the leaf nodes show the classification records. At each node, the data is split using a split criterion, and this process is repeated until the leaf node is reached. This study uses DT with two hyperparameters maximum depth and random state, which are set as 150 each. The max_depth hyperparameter limits the decision trees to a maximum 150 level. When we set the max depth in the range of 10 to 100, it becomes too simple and unable to detect noise or irrelevant features from the large dataset, resulting in poor generalization performance on new or unseen data. We can manage the complexity of the tree and minimize the possibility of overfitting by setting the maximum depth to 150. Also, the random state value 150 was not commonly used in the previous research; we experimented with setting it to 150 to maintain consistency and handle the order of random numbers.

K Nearest Neighbors. KNN classifier is employed to solve the regression and classification problems [11]. KNN is called the lazy learners, as it uses all the data for training and new samples are classified based on similarity. The similarity calculation is done using distance measurement between the new samples and existing class samples and metrics such as Euclidean, Miknowski, and so on, are used.

Long Short-term Memory. LSTM is an **artificial neural network (ANN)** that is applied for classification problems [33]. LSTM has feedback connections. LSTM can process entire sequences of data. This distinguishing functionality makes LSTM models ideal to process and predicting the data. Moreover, LSTM is used for several tasks, including robot control, video games, speech recognition, connected hand-writing recognition, machine translation, and healthcare. LSTM is capable to forget information or recall information due to its forget gate. It has an output gate, forget gate, update gate, and input gate. The forget gate is used to determine which information is thrown away from the cell state

$$T_F = \phi(S_R.[D_F - 1, P_F] + Y_R), \tag{7}$$

where D_F is used for the weighted matrix, and Y_R is for the bias vector.

Consider T_F is a number from 0 and 1 where 0 is used to forget the value and 1 is used to keep the value.

$$A_F = \phi(S_Q.[D_F - 1, A_F] + Y_F), \tag{8}$$

$$G_F = tanh(S_G.[D_F - 1, A] + Y_W), \tag{9}$$

where S_Q and S_G represent weighted matrices, and Y_F and Y_W show the bias vectors.

For the output, A_F and G_F are described.

$$G_F = T_F * W_{F-1} + A_F * W_F, \tag{10}$$

where T_F is used for forgetting information. Also, $A_F G_F$ selects the total number of values that are for the modification of the cell.

$$A_E = \phi(Y_E.[D_F - 1, A_F] + Y_E), \tag{11}$$

$$H_F = E_F * tanh(f_F), \tag{12}$$

where A_E is used for the output state. The new cell state G_F is multiplied withE_F. Moreover, the $tanh$ function is to achieve H_F, which is the output state of A_E.

Gated Recurrent Unit. The GRU is a recent generation of RNN and works similarly to LSTM. GRU employs hidden states to transfer information. The update gate and a reset gate are the two main gates of GRU [28]. The working of the update gate is similar to the input and the forget gate of an LSTM. The update gate is used to decide which information is to be thrown away and which information should be kept. However, the reset gate decides how much information needs to be forgotten from the previous information.

Recurrent Neural Network. RNN model is a class of ANNs used for classification [42].

RNN uses the internal memory to process the variable length classifications of inputs. RNN has been used for different tasks such as connected hand-writing recognition, speech recognition, text analysis, and so on. RNN is ideally Turing complete and may run random programs to further process random input sequences. Due to the internal memory, RNN remembers the previous inputs. RNN simulates a discrete-time dynamic behavior system that has s_i for the input layer, f_i for the hidden layer, and h_i for the output layer, while i is used to denote time. The dynamical model is defined as

$$f_i = R(s_i, f_{i-1}), \tag{13}$$

Table 5. Architecture and Fine-tuned Hyperparameters of Deep Learning Models

Model	Trainable Parameters	
CNN	524,195	Embedding layer = (5,000*100) Loss = categorical_crossentropy Optimizer = Adam, Epochs = 25, Batch_size = 64
LSTM	624,819	
BILSTM	626,787	
RNN	533,539	
GRU	592,547	

$$h_0 = R0(f_i), \tag{14}$$

where Rf and $R0$ are described as functions for the state transition and output, respectively.

The above-discussed deep learning models are optimized regarding the structure and hyperparameters and a complete list of such parameters is given in Table 5.

Bidirectional Encoder Representations from Transformers. BERT model is based on the transformer structure [34], especially, BERT contains transformer encoder layers. BERT is a pre-trained model using language representations, and it has been trained on a large text like Wikipedia. It can be then applied to other NLP tasks, including sentiment analysis and question-answering. The model is conceptually easy and empirically more powerful. The BERT model can be fine-tuned using a few resources like smaller datasets to optimize its performance. Since the pre-training stage involves extensive computation, fine-tuning part requires fewer computational resources.

Distilled Version of BERT. This model has a similar general architecture to BERT. DistilBERT model is a fast, small, and light transformer model that is based on the BERT model [44]. Knowledge distillation is achieved during the pre-training stage to decrease the size of a BERT system by 40%, despite the reduction in size, the system still retains 97% of its language understanding capabilities and becomes 60% faster in processing. The researchers introduced a triple loss merging distillation, language modeling, and cosine-distance losses.

XLNet. XLNet is another well-known and recent model to carry out NLP tasks [43]. The XLNet combines the latest advances in NLP using innovative choices on how to solve language modeling issues. Moreover, the XLNet is trained on a huge corpus **general language understanding evaluation (GLUE)** benchmark and can attain state-of-the-art performance for standard NLP tasks. A modified language model training objective is the main contribution of the XLNet model that is used to learn the conditional distributions from all possible permutations for the tokens in order.

Robustly Optimized RoBERT Pretrained Approach. RoBERTa is an optimized model on the concept of BERT by modifying the static-mask to a dynamic-mask [37]. RoBERTa model has increased the input text encoding and is trained with large batch sizes. Dynamic masking is used to predict masked tokens with various probabilities in RoBERTa. However, BERT applies static masking where similar tokens are masked with a similar probability. Additionally, RoBERTa is used for more aggressive data methods, including back translation and sentence breaking and to improve the size of training data. RoBERTa is trained on more diverse and large-scale datasets that contain a wide range of text types such as scientific articles and web pages. The parameters for Transformer-based models are presented in Table 6.

3.5 Building Hybrid Transformer-based Models

BERT and RoBERTa models are transformer-based models, both developed by Google and Facebook AI research groups, respectively. The hybrid models are designed to address complex

Table 6. Parameters and Layers Used for Transformer-based Models

RoBERTa Model	**Parameters**	**XLNET Model**	**Parameters**	**DistilBERT Model**	**Parameters**
input_ids	0	input_ids	0	input_ids	0
RoBERTa main layer	124,645,632	TF XLNet Main Layer	116,718,336	TFDistilBert Main layer	66,362,880
Dropout	0	TF operators	0	TF operators	0
Dense	2,307	TFOpLambda	0	Dropout	0
Trainable Parameters	124,647,939	Dropout	0	Dense	2,307
		Dense	2,307	Trainable Parameters	66,365,187
		Trainable Parameters	116,720,643		
BERT Model	**Parameters**	**RoBERTa+BERT**	**Parameters**	**RoBERTa+DistilBERT**	**Parameters**
input_ids	0	input_ids	0	input_ids	
TF BERT Main Layer	109,482,240	TF RoBERTa Main Layer	124,645,632	TF RoBERTa Main Layer	124,645,632
Dropout	0	TF BERT Main Layer	109,482,240	TF DistilBERT Main Layer	66,362,880
Dense	49,216	Dropout	0	TF Operators	0
Dense	2,080	Pooled output	0	Dropout	0
Dropout	0	TF.Concate	0	Pooled output	0
Dense	99	Dense	4,611	TF.Concate	0
Trainable Parameters	109,533,635	Trainable Parameters	234,132,483	Dense	4,611
				Trainable Parameters	191,013,123

problems and improve the classification results in many NLP tasks [21]. Hybrid models are more robust, because they are trained on structured as well as unstructured data and perform well on unseen data. Hybrid models enhance the results of transformer models as compared to single models.

Hybrid models are built using TensorFlow, Keras, and Transformers libraries and modules. For this purpose, transformer pre-trained models are loaded, including RoBERTa and BERT from transformer libraries. Appropriate tokenizers are used for RoBERTa and BERT. Then an embedding layer is created for hybrid models using the functional Keras **application programming interface (API)**. To prevent overfitting and represent the input sequence as a fixed size, dropout and pooled output layers are used, respectively. After that, the pooled output layers of both models are concatenated using the "tf.concat" layer. Finally, a dense classification layer is used to classify the sentiments. For compilation, the "Adam" optimizer is used and the categorical cross-entropy is used as the loss function. The architecture of the proposed hybrid models is given in Figure 2.

3.6 Performance Metrics

Performance metrics such as precision, recall, F1 score, and accuracy are used to test the performance of a model. The model's evaluation is very critical for getting quality results. To assess the performance of evaluation metrics, **true positive (TP), true negative (TN), false positive (FP)**, and **false negative (FN)** rates are used. TP means the model predicts positive class accurately, while TN means the model predicts accurately negative class. FP indicates a positive class predicted incorrectly, whereas FN means a negative class predicted incorrectly. The highest accuracy rate is 1, and the lowest accuracy rate is 0. The following equations are used to calculate performance evaluation metrics:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}, \tag{15}$$

$$Precision = \frac{TP}{(TP + FP)}, \tag{16}$$

$$Recall = \frac{TP}{(TP + FN)}, \tag{17}$$

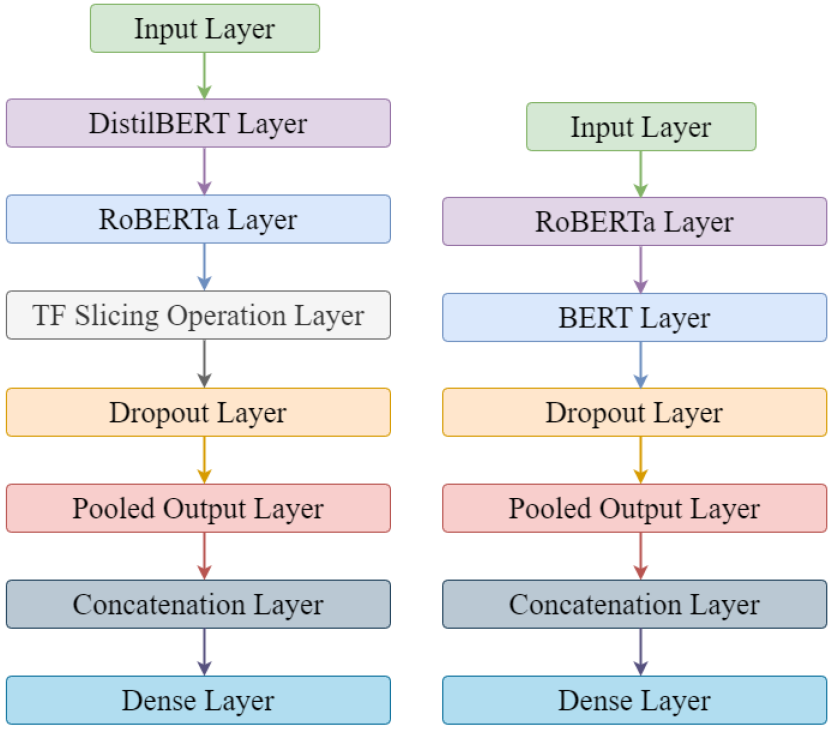

Fig. 2. Architecture of proposed hybrid models, **(a)** DistilBERT+RoBERTa and **(b)** RoBERTa+BERT hybrid model.

$$F1 - score = 2 \times \frac{(Recall \times precision)}{(Recall + precision)}. \tag{18}$$

4 RESULTS AND DISCUSSION

Experimental results of the machine and deep learning models are presented in this section with BoW and TFIDF features. Also, results from four transformer-based and two hybrid models are presented for sentiment analysis on ChatGPT Arabic tweets.

4.1 Results with BoW Features

Table 7 depicts the results of machine learning models for Arabic tweet sentiment analysis. Accuracy, precision, recall, and F1 score are metrics to assess the results of various models. The macro-average of all classes is also presented. The LR model attained a 93.84% accuracy and 94% macro average, which is the highest among all machine learning models. Its performance is followed by the KNN, which obtains an 89.38% average accuracy for positive, negative, and neutral classes. The poorest performing model is DT, which attains a 79.80% accuracy.

4.2 Results Using TFIDF Features

Sentiment analysis is also performed using TFIDF features with machine learning models. Table 8 shows that KNN obtains the highest accuracy of 89.81% using TFIDF features, which are followed by LR with an 87.81% accuracy. Apart from marginal improvement in KNN, the performance of

Table 7. Performance of Machine Learning Models Using BoW Features

Model		Precision	Recall	F1 score	Accuracy
LR	positive	93	93	93	93.84
	negative	96	99	98	
	neutral	93	89	91	
	macro avg	94	94	94	
DT	positive	90	67	70	79.80
	negative	93	86	89	
	neutral	65	86	74	
	macro avg	83	80	80	
RF	positive	92	70	79	78.90
	negative	93	77	84	
	neutral	64	90	74	
	macro avg	83	79	79	
KNN	positive	92	81	86	89.38
	negative	98	96	97	
	neutral	80	91	85	
	macro avg	90	89	89	
SVM	positive	82	81	81	82.41
	negative	82	91	86	
	neutral	83	76	79	
	macro avg	82	82	82	

other models is degraded when used with TFIDF features, as models perform better when used with BoW features.

Figure 3(a) illustrates the correct and wrong predictions for machine learning models using BoW features. Using BoW features, LR made 5,214 correct predictions, which is higher than other models, and only 342 wrong predictions out of 5,556. RF model achieved the highest number of wrong predictions and the lowest number of correct predictions. DT model made 1,122 wrong and 4,434 correct predictions. Figure 3(b) shows that both LR and KNN models make the highest correct and lowest wrong predictions. DT model also performs poorly with the lowest correct predictions using TFIDF features.

4.3 Results of Deep Learning Models

The results of Arabic tweet sentiment analysis using deep learning models are presented in Table 9. Table 9 shows that LSTM obtains the best performance for Arabic tweets classification with a 94.26%. LSTM is utilized with two layers utilizing 100 and 64 units, one dense 32-unit layer, and one classification layer. CNN shows a slightly lower performance with an accuracy of 94.06% while RNN, BiLSTM, and GRU also perform better. Results indicate that deep learning deals efficiently with large data and performs well. Deep learning models learn complex structures effectively. All models performed well on the Arabic tweet classification.

Table 8. Performance of Machine Learning Models Using TFIDF Features

Model		Precision	Recall	F1 score	Accuracy
LR	positive	89	86	87	87.81
	negative	91	95	93	
	neutral	84	82	83	
	macro avg	88	88	88	
DT	positive	91	61	73	76.92
	negative	93	81	87	
	neutral	60	88	72	
	macro avg	82	77	77	
RF	positive	92	72	81	80.50
	negative	92	82	87	
	neutral	66	87	75	
	macro avg	83	81	81	
KNN	positive	96	79	86	89.81
	negative	99	94	97	
	neutral	78	96	86	
	macro avg	91	90	90	
SVM	positive	84	77	80	83.98
	negative	88	99	93	
	neutral	79	75	77	
	macro avg	84	84	84	

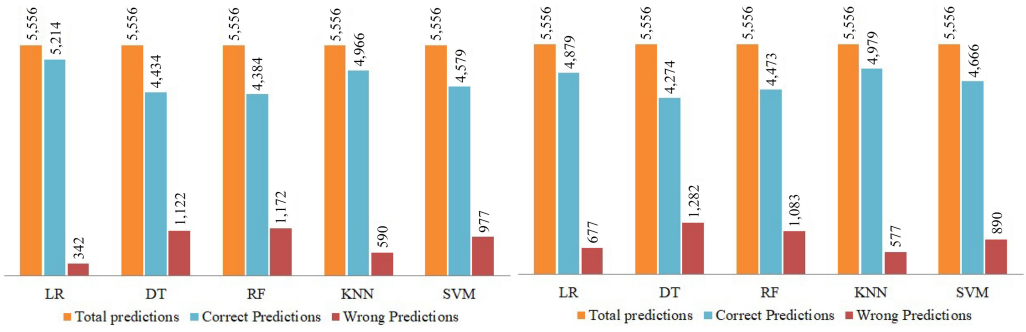

Fig. 3. Graphical representation of confusion matrices for machine learning models using **(a)** BoW features and **(b)** TFIDF features.

The predictions of deep learning models are represented in Figure 4. Figure 4 shows that the LSTM model makes 5,237 correct predictions, which is the highest among all deep learning models while the total wrong predictions are 319, which is the lowest compared to other deep learning predictions. The highest number of wrong predictions is made by the RNN model, which is 367.

Table 9. Performance of Deep Learning Model for Arabic Tweets

Model		Precision	Recall	F1 score	Accuracy
CNN	positive	93	94	94	94.06
	negative	97	98	97	
	neutral	92	90	91	
	macro avg	94	94	94	
LSTM	positive	94	95	95	94.26
	negative	96	97	97	
	neutral	92	91	91	
	macro avg	94	94	94	
RNN	positive	91	96	93	93.39
	negative	96	98	97	
	neutral	94	87	90	
	macro avg	93	93	93	
BiLSTM	positive	93	94	93	93.48
	negative	97	97	97	
	neutral	91	90	90	
	macro avg	93	93	93	
GRU	positive	93	95	94	93.59
	negative	95	98	96	
	neutral	93	88	90	
	macro avg	94	94	94	

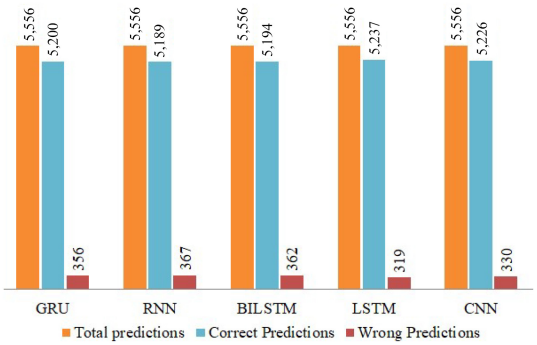

Fig. 4. Graphical representation of confusion matrices for deep learning models.

4.4 Results of Transformer-based Models

The results of transformer-based models are represented in Table 10. The RoBERTa model achieved an overall accuracy of 84.18%, while the achieved precision, recall, and the F1 score are 95%, 90%, and 92%, respectively. The DistilBERT model achieved better accuracy than RoBERTa and XL-Net with an 85.08% accuracy, which is the best among all transformer models. The XLNet model

Table 10. Results of Transformer-based Models

Model		Precision	Recall	F1 score	Accuracy
RoBERTa	positive	87	76	82	84.07
	negative	97	89	93	
	neutral	68	88	77	
	macro avg	84	85	84	
DistillBERT	positive	86	83	85	85.08
	negative	98	85	91	
	neutral	71	88	79	
	macro avg	85	85	85	
XLNet	positive	60	58	59	65.17
	negative	78	70	73	
	neutral	58	67	62	
	macro avg	65	65	65	
BERT	positive	85	78	81	83.59
	negative	94	87	90	
	neutral	72	87	79	
	macro avg	84	84	83	

Table 11. Results of Transformer-based Hybrid Models

Model		Precision	Recall	F1 score	Accuracy
RoBERTa + DistilBERT	positive	99	93	96	95.97
	negative	100	96	97	
	neutral	89	98	93	
	macro avg	96	96	96	
RoBERTa + BERT	positive	99	93	96	96.02
	negative	100	97	98	
	neutral	89	99	94	
	macro avg	96	96	96	

achieved an average accuracy of 65.17%, while the macro average of precision, recall, and F1 score are low for the XLNet model. The XLNet model does not perform well.

4.5 Results of Proposed Hybrid Transformer-based Models

Table 11 shows the performance analysis of hybrid models. The hybrid model RoBERTa+ DistilBERT achieved an accuracy of 95.97% and a precision of 100% precision for the negative class. The RoBERTa+BERT hybrid model performs very well, achieving a 96.02% accuracy with a 96% macro average for precision, recall, and F1 score. Its performance is the best among all the machine learning, deep learning, and transformer-based models employed in this study.

Figure 5(a) shows that the DistillBERT model made 829 wrong predictions, which is the lowest in terms of other transformers. The XLNet model attained the highest wrong predictions. Figure 5(b) shows the confusion matrix for the proposed hybrid transformer-based models. It indicates the significance of the proposed model in terms of correct predictions. The proposed hybrid model RoBERTa+BERT achieved 5,335 correct predictions and 221 wrong predictions. Also, the RoBERTA+DistilBERT hybrid model achieved 5,332 correct and 224 wrong predictions. Overall, hybrid transformer-based models achieved the best results.

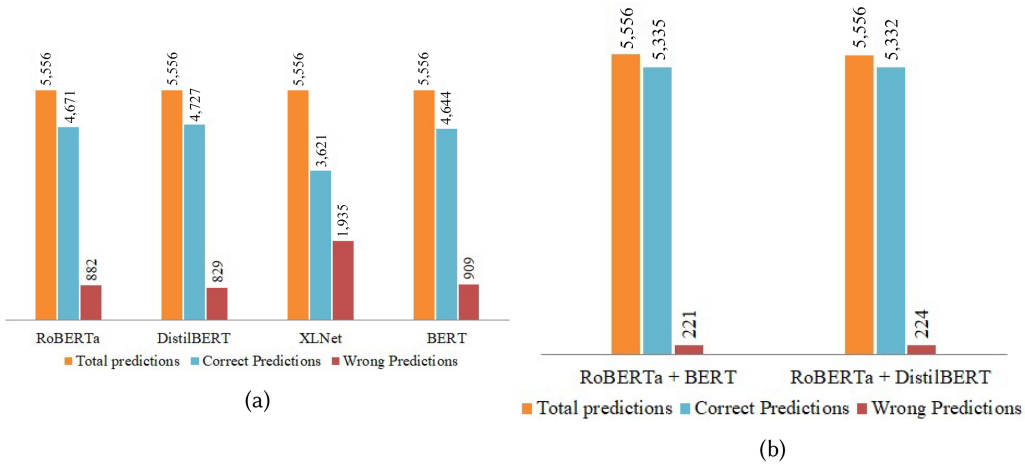

Fig. 5. Graphical representation of confusion matrices for **(a)** Transformer-based models and **(b)** Proposed hybrid model.

4.6 Training and Testing Curves of Hybrid Transformer-based Models

Training accuracy is used to assess the training performance of a model, and testing accuracy indicates how well the model behaves on testing data after training is done. Training loss is used to measure the learning process of a model on training data, and testing loss, which is an efficient evaluation metric, indicates how well a model performs on unseen data. In this study, the categorical cross-entropy loss function is used to train the model for classification. Figure 6(a) shows the training and testing accuracy of the RoBERTa+DistilBERT model, where at epoch 1, the training accuracy is 0.6330, and at epoch 10, it is 98.62. At epoch 7, testing accuracy is 96.08.

Figure 6(b) shows the training loss is at its highest of 0.7719 at epoch 1, and the testing loss is 0.5018. The testing loss is lowest at epoch 5, and at epoch 10, the loss reached 0.1742. Figure 6(c) shows the training and testing accuracy of the RoBERTa+BERT model, in which training accuracy is lowest at epoch 1 and after that, it increases along with testing accuracy. Figure 6(d) shows the training and testing loss of the RoBERTa+BERT model.

4.7 Comparison of Hybrid Transformer Models with State-of-the-art Existing Models

Hybrid transformer-based models are compared with existing state-of-the-art studies in the literature to prove the significance of the proposed approach. Table 12 illustrates the experimental results of the proposed hybrid model with state-of-the-art sentiment analysis studies. All experiments described in the literature in Table 12 are performed on Twitter datasets. For example, Alqarni et al. [6] employed Arabic Twitter data for sentiment analysis using deep learning and attained a 92.8% accuracy with CNN. Similarly, Hassan et al. [2] used a hybrid of CNN and LSTM to detect hate-speech from the Arabic data with a 75% accuracy. Bayati et al. [1] took an Arabic book-review dataset in 2020 and applied LSTM deep learning model for sentiment analysis tasks and reported an accuracy of 82%. Furthermore, Cheng al et al. [15] and Salaman et al. [7] utilized the Twitter dataset for sentiment analysis; however, the obtained results are not very promising. A study conducted by H. Chouikhi et al. [16] on sentiment analysis of Arabic tweets using an optimized BERT model achieved 91% accuracy with the proposed Arabic BERT model. However, other metrics are not utilized to assess the performance of the proposed model, and the accuracy achieved is not exceptional. Fsih et al. [25] utilized BERT model for Arabic sentiment classification. They used a limited dataset and applied augmentation to enhance the dataset samples to

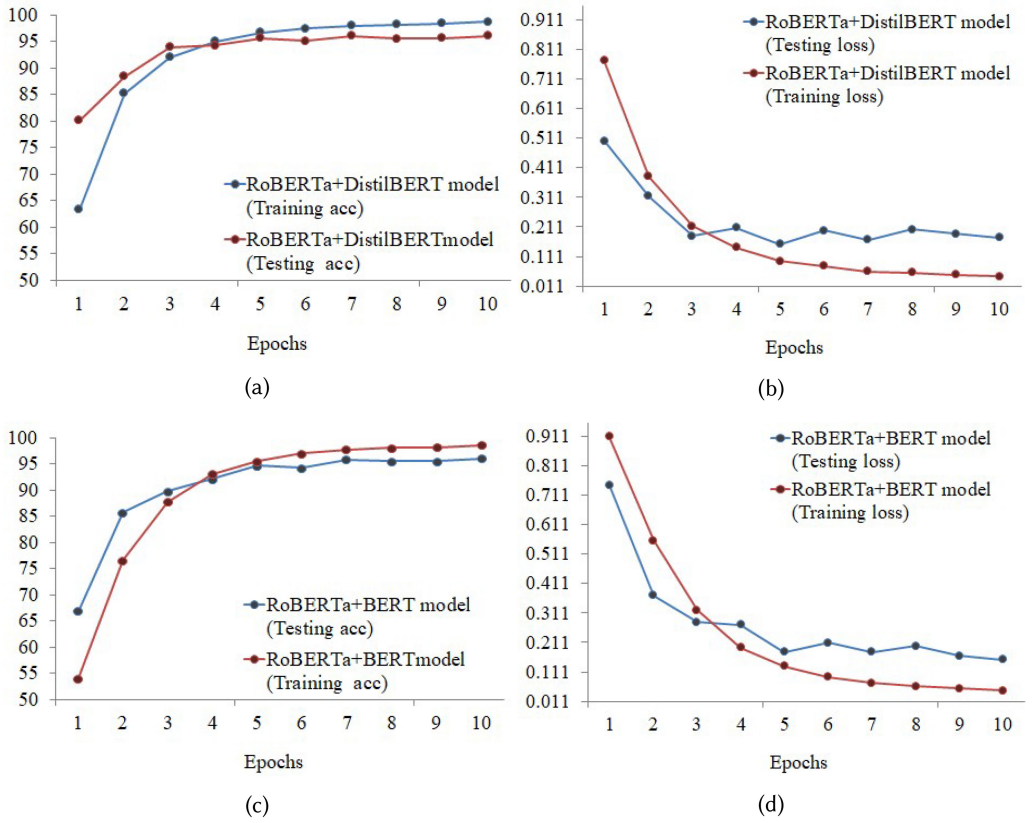

Fig. 6. Training and testing curves of hybrid transformer-based models: **(a)** Training and testing accuracy of RoBERTa + DistilBERT model, **(b)** Training and testing loss of RoBERTa + DistilBERT model, **(c)** Training and testing accuracy of RoBERTa + BERT model, and **(d)** Training and testing loss of RoBERTa + BERT model.

improve the classification performance. With augmented data, authors achieved a 67% improved F1 score for Arabic tweets. Fawzy et al. [24] used the last four fine-tuned layers of the BERT model, while others remained frozen and were combined with the CNN classification model, to predict the sentiments using low-resource Arabic tweets. They performed experiments using three Arabic Twitter datasets and achieved the highest accuracy of 94%. Another study conducted by Khered et al. [35] used ensemble MARBERT with hyperparameter optimizations. The dataset is split into training, development, and testing phases. They ensemble the models in different ways, but how to ensemble them in different ways is not well explored. Also, the weighted average accuracy score was achieved by the Ensemble MARBERT model at 74%. Performance comparison in Table 12 indicates that the proposed hybrid transformer-based model achieves the highest accuracy compared to previous state-of-the-art Arabic tweets sentiment analysis studies.

4.8 Cross-validation Results

We performed cross-validation experiments on ML, DL, and proposed transformer-based models to evaluate their significance. Table 13 presents the cross-validation accuracy score as well as the standard deviation. LR achieved the highest cross-validation accuracy of 93.7% and +/− 0.015 standard deviation score. The deep learning LSTM model achieved a 93.12% validation score and +/− 2.43 standard deviation. The Ensemble RoBRTa+BERT achieved excellent 95.64% validation

Table 12. Comparison of Hybrid Models with State-of-the art Models

Paper Ref.	Methods	Dataset used	Accuracy	Published
[15]	LSTM	Arabic Twitter data	88%	2019
[1]	LSTM	Arabic book-reviews	82%	2020
[7]	DMNB	Arabic Twitter data	87.5%	2020
[2]	Hybrid CNN+LSTM	Arabic Twitter data	75%	2021
[6]	CNN	Arabic Twitter data	92.8%	2023
[16]	Optimized BERT	Arabic Twitter data	91%	2021
[25]	BERT	Arabic Twitter data	67%	2022
[24]	BERT+CNN	Arabic Twitter data	94%	2022
[35]	MARBERT	Arabic Twitter data	74%	2022
This study	**Hybrid RoBERTa+BERT**	**Arabic Twitter data**	**96.02%**	**2023**

Table 13. Cross-validation Results

Model	Accuracy (%)	Standard Deviation (SD)
LR	93.7	±0.015
DT	79.6	±0.012
RF	77.5	±0.028
KNN	91.01	±0.012
SVM	81.05	±0.022
LSTM	93.12	±2.43
RNN	93.74	±1.14
BILSTM	92.23	±2.42
CNN	93.73	±1.31
GRU	92.79	±1.35
BERT	0.82	±0.01
RoBERTa+BERT	95.64	±0.33
RoBERTa+DistilBERT	95.52	±0.37

accuracy with +/− 0.33 standard deviation. The cross-validation results proved that ensemble transformer-based models yielded outclass results and showed their significance in Arabic tweets classification.

4.9 Discussion

Table 7 demonstrates that, using BoW features, LR achieved the maximum overall accuracy (95%) and RF achieved the lowest (79%) while using TFIDF features; KNN obtained 90% and DT 77% accuracy overall. The DL models, including LSTM and RNN, attained an overall accuracy of 94% and 93%, respectively. The single transformer-based models obtained an overall maximum accuracy of 85% and minimum accuracy of 65%. The hybrid transformer-based RoBERTa+BERT model proposed achieved an overall accuracy of 96%, while RoBERTa+ DistilBERT achieved an overall accuracy of 95%. Addressing bias and generalizations, hybrid models perform extremely well in terms of overall accuracy and precision score. The model with the most correct predictions is the hybrid of RoBERTa and BERT, which makes 5,335 (96%) correct predictions overall, whereas with a total of 1,935 incorrect predictions on the test data, the XLNET model stands out as the model with the highest number of wrong predictions. Using TFIDF features, DT makes 1,282 incorrect predictions in ML models, while RNN makes 367 incorrect predictions in DL models. Finally,

Table 14. Statistical T-test Comparison

Comparison Case	Independent T-test		Paired T-test	
	P-Value	T	P-Value	T
RoBERTa + BERT vs. RoBERTa	5.51e-09	47.9	2.02e-05	47.6
RoBERTa + BERT vs. DSTILBERT	2.94e-15	532.8	5.61e-09	732.3
RoBERTa + BERT vs. DSTILBERT	4.83e-09	49.0	1.83e-05	49.3
RoBERTa + BERT vs. CNN	1.69e-11	125.8	2.79e-07	198.9
RoBERTa + BERT vs. LSTM	9.54e-08	29.7	6.50e-05	32.3
RoBERTa + BERT vs. GRU	8.67e-07	20.5	0.00028	19.6
RoBERTa + BERT vs. RNN	9.50e-08	29.7	7.07e-05	31.4
RoBERTa + BERT vs. LR	1.82e-09	57.6	1.66e-05	51.0
RoBERTa + BERT vs. KNN	1.31e-07	28.2	9.74e-05	28.2

results proved that the proposed Hybrid model achieved the most correct predictions in terms of four performance metrics as compared to other ML, DL, and single transformer-based models.

Transformer-based hybrid models are more robust and reliable in handling noise in the data, outliers, and generalizations due to large training data, self-attention mechanisms, and the utilization of relevant global contextual information. However, ML models do not handle noisy data and outliers due to their poor generalization performance and hand-crafted features. In contrast to ML, DL models are better able to handle noisy input and learn complicated hierarchical structures.

According to the experimental results, neutral tweets are more challenging to classify than positive and negative tweets. The neutral tweets provide general information about any context but do not provide strong and clear positive or negative statements. Curing and filtering neutral tweets may be challenging. Proposed transformer-based hybrid models are able to capture nuances in the language as well as the context to improve the classification accuracy of neutral tweets.

The positive sentiments provide positive opinions, satisfaction, happiness, or preferences about ChatGPT. Positive sentiments about ChatGPT technology are very helpful to improve its model behavior, building strong questioning-and-answering Chabot, knowing their strengths, enhancing the quality of service and meeting the user's requirements, and providing 24-hour quick service with flexibility.

Table 14 shows the statistical test performed on Arabic tweet classification. A statistical test is performed using two types of tests, i-e paired T-test and the independent T-test. We compared our proposed approach with other approaches to prove its significance. A small p-value demonstrates that there is very solid evidence against the null hypothesis. If the p-value is less than the significance level (p-value < 0.05), then it indicates that the observed difference is statistically important and the Null Hypothesis can be rejected. In our case, paired T-test results achieved 2.02e-05 smallest p-value, indicating a significant difference between the two RoBERTa + BERT vs. RoBERTa models. The proposed method outperforms others by a substantial margin.

4.10 Challenges and Limitations of ChatGPT

The most powerful AI-based model, ChatGPT, is trained on massive data sets using a transformer-based model to generate human-like text. ChatGPT offers multiple benefits, minimizes the demand for customer service, and efficiently satisfies customer needs. However, the ChatGPT model has some limitations that produce erroneous outputs:

- ChatGPT generates incorrect or inaccurate responses with important information missing.
- ChatGPT provides nonsensical replies to specific queries.

- ChatGPT cannot comprehend extremely intricate sentences and is emotionless in certain circumstances.
- Another limitation is that it is unsuitable for lengthy content or provides an incomplete response for lengthy content on time.
- Additionally, it sometimes provides incorrect references to paragraphs that are unavailable across Google platforms. Additionally, it cannot perform multiple tasks simultaneously.

ChatGPT does not produce or generate creative or original real-world information due to offensive or harmful language biases. It is very challenging when generated responses through ChatGPT are not according to the user's expectations or generate information that is not related to the topic of the user's input. ChatGPT cannot create up-to-date information and cannot respond to current scenarios. The biggest challenge is potential security concerns for the ChatGPT model: Hackers deploy the model architecture with malicious involvements that produce invalid outputs.

4.11 Limitations of Study

This study uses the TextBlob approach for sentiment analysis, which may not be suitable for all types of text data and languages. It is a lexicon approach that employs pre-built lexicons for sentiment analysis, thus unable to capture nuances. Furthermore, TextBlob ignores unknown words and considers only those words to whom it can assign a polarity score. Sentences containing irony and sarcasm are ambiguous and may exhibit multipolarity. That is the reason TextBlob does not yield the best results if the text contains multiple languages.

Valency aware dictionary for sentiment reasoning (VADER) is another popular approach for sentiment analysis that we intend to explore. Although it may face similar problems, it has more emphasis on social media and can produce good results for texts containing emojis, punctuation, and so on.

We intend to utilize embedding-based models for our future work like Flair, which is a pre-trained model. Unlike TextBlob and VADER, which are rule-based approaches that ignore the context of a whole sentence, Flair can determine the sentiment of a sentence. Although Flair is much slower as compared to rule-based approaches, it can provide higher accuracy.

5 CONCLUSIONS

Sentiment analysis of tweets provides valuable feedback from the users to improve the quality of products and services. This study provides a hybrid model for analyzing tweets regarding ChatGPT, which gained remarkable attention lately. However, unlike English tweets for which a large body of works already exists, this study uses Arabic tweets, which is very less studied, as do the transformer-based models for sentiment analysis. The proposed hybrid model comprises RoBERTa and BERT, and extensive experiments are performed to evaluate its performance against machine learning and deep learning models, as well as transformer-based models and existing state-of-the-art models for Arabic tweets analysis. Results indicate superior performance of the proposed approach compared to other models with an accuracy of 96.02%. Experiments show that hybrid transformer-based achieved far better results than other single transformer and machine learning models. In the future, we intend to increase the size of the gathered dataset and utilize advanced hybrid approaches to further improve the accuracy of sentiment analysis.

REFERENCES

[1] Abdulhakeem Qusay Al-Bayati, Ahmed S. Al-Araji, and Saman Hameed Ameen. 2020. Arabic sentiment analysis (ASA) using deep learning approach. *J. Eng.* 26, 6 (2020), 85–93.

[2] Areej Al-Hassan and Hmood Al-Dossari. 2022. Detection of hate speech in Arabic tweets using deep learning. *Multimedia Systems* 28 (2022), 1963–1974. https://doi.org/10.1007/s00530-020-00742-w

[3] Arwa A. Al Shamsi and Sherief Abdallah. 2021. Text mining techniques for sentiment analysis of Arabic dialects: Literature review. *Adv. Sci. Technol. Eng. Syst. J.* 6 (2021), 1012–1023.
[4] Haifa K. Aldayel and Aqil M. Azmi. 2016. Arabic tweets sentiment analysis–A hybrid scheme. *J. Inf. Sci.* 42, 6 (2016), 782–797.
[5] Alanoud Mohammed Alduailaj and Aymen Belghith. 2023. Detecting Arabic cyberbullying tweets using machine learning. *Mach. Learn. Knowl. Extract.* 5, 1 (2023), 29–42.
[6] Arwa Alqarni and Atta Rahman. 2023. Arabic tweets-based sentiment analysis to investigate the impact of COVID-19 in KSA: A deep learning approach. *Big Data Cogn. Comput.* 7, 1 (2023), 16.
[7] Hussain AlSalman. 2020. An improved approach for sentiment analysis of Arabic tweets in Twitter social media. In *Proceedings of the 3rd International Conference on Computer Applications & Information Security (ICCAIS'20)*. IEEE, 1–4.
[8] Kushal Anjaria. 2020. Negation and entropy: Effectual knowledge management equipment for learning organizations. *Expert Syst. Applic.* 157 (2020), 113497.
[9] Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraGPT2: Pre-trained transformer for Arabic language generation. *arXiv preprint arXiv:2012.15520* (2020).
[10] Gérard Biau and Erwan Scornet. 2016. Rejoinder on: A random forest guided tour. *Test* 25 (2016), 264–268.
[11] Vishwanath Bijalwan, Vinay Kumar, Pinki Kumari, and Jordan Pascual. 2014. KNN based machine learning approach for text and document mining. *Int. J. Datab. Theor. Applic.* 7, 1 (2014), 61–70.
[12] Som Biswas. 2023. ChatGPT and the future of medical writing. *Radiology* 307, 2 (2023), e223312. DOI: 10.1148/radiol.223312
[13] Xavier A. Carrasco, Ashraf Elnagar, and Mohammed Lataifeh. 2021. A generative adversarial network for data augmentation: The case of Arabic regional dialects. *Procedia Comput Sci.* 189 (2021), 92–99.
[14] Birol Çelik, Hüseyin Uzunboylu, and Nur Demirbaş-Çelik. 2023. Higher education students' social media platform preferences for educational purposes. *Revista de Educación a Distancia (RED)* 23, 72 (2023).
[15] Li-Chen Cheng and Song-Lin Tsai. 2019. Deep learning for automated sentiment analysis of social media. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 1001–1004.
[16] Hasna Chouikhi, Hamza Chniter, and Fethi Jarray. 2021. Arabic sentiment analysis using BERT model. In *Proceedings of the 13th International Conference on Advances in Computational Collective Intelligence*. Springer, 621–632.
[17] K. R. Chowdhary. 2020. *Fundamentals of Artificial Intelligence*. Springer.
[18] Jingfeng Cui, Zhaoxia Wang, Seng-Beng Ho, and Erik Cambria. 2023. Survey on sentiment analysis: Evolution of research methods and topics. *Artif Intell Rev* 56 (2023), 8469–8510. https://doi.org/10.1007/s10462-022-10386-z
[19] Abdelhalim Hafedh Dahou and Mohamed Amine Cheragui. 2023. Named entity recognition for Algerian Arabic dialect in social media. In *Proceedings of the 12th International Conference on Information Systems and Advanced Technologies (ICISAT 2022): Intelligent Information, Data Science and Decision Support System*. Springer, 135–145.
[20] Robert Dale. 2021. GPT-3: What's it good for? *Nat. Lang. Eng.* 27, 1 (2021), 113–118.
[21] Cach N. Dang, Maria N. Moreno-Garcia, and Fernando De la Prieta. 2021. Hybrid deep learning models for sentiment analysis. *Complexity* 2021 (2021), 1–16.
[22] Bijoyan Das and Sarit Chakraborty. 2018. An improved text sentiment classification model using TF-IDF and next word negation. *arXiv preprint arXiv:1806.06407* (2018).
[23] Ibrahim Abu Farha and Walid Magdy. 2022. The effect of Arabic dialect familiarity on data annotation. In *Proceedings of the 7th Arabic Natural Language Processing Workshop (WANLP'22)*. 399–408.
[24] Mohamed Fawzy, Mohamed W. Fakhr, and Mohamed Abo Rizka. 2022. Sentiment analysis for Arabic low resource data using BERT-CNN. In *Proceedings of the 20th International Conference on Language Engineering (ESOLEC'22)*. IEEE, 24–26.
[25] Emna Fsih, Saméh Kchaou, Rahma Boujelbane, and Lamia Hadrich Belguith. 2022. Benchmarking transfer learning approaches for sentiment analysis of Arabic dialect. In *Proceedings of the 7th Arabic Natural Language Processing Workshop (WANLP'22)*. 431–435.
[26] Catherine A. Gao, Frederick M. Howard, Nikolay S. Markov, Emma C. Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T. Pearson. 2022. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *bioRxiv* (2022), 2022–12.
[27] Emitza Guzman and Walid Maalej. 2014. How do users like this feature? A fine grained sentiment analysis of app reviews. In *Proceedings of the IEEE 22nd International Requirements Engineering Conference (RE'14)*. IEEE, 153–162.
[28] Haret Hadhood. 2022. *Stock Trend Prediction Using Deep Learning Models LSTM and GRU with Non-linear Regression*. Master's thesis. Itä-Suomen yliopisto.
[29] Mohammed Hadwan, Mohammed Al-Sarem, Faisal Saeed, and Mohammed A. Al-Hagery. 2022. An improved sentiment classification approach for measuring user satisfaction toward governmental services' mobile apps using machine learning methods with feature engineering and SMOTE technique. *Appl. Sci.* 12, 11 (2022), 5547.

[30] Ahmad S. Haider and Riyad F. Hussein. 2022. Modern standard Arabic as a means of euphemism: A case study of the MSA intralingual subtitling of Jinn series. *J. Intercult. Commun. Res.* 51, 6 (2022), 628–643.
[31] Mubin Ul Haque, Isuru Dharmadasa, Zarrin Tasnim Sworna, Roshan Namal Rajapakse, and Hussain Ahmad. 2022. "I think this is the most disruptive technology" : Exploring sentiments of ChatGPT early adopters using Twitter data. *arXiv preprint arXiv:2212.05856* (2022).
[32] Maha Heikal, Marwan Torki, and Nagwa El-Makky. 2018. Sentiment analysis of Arabic tweets using deep learning. *Procedia Comput. Sci.* 142 (2018), 114–122.
[33] Xin Hong, Rongjie Lin, Chenhui Yang, Nianyin Zeng, Chunting Cai, Jin Gou, and Jane Yang. 2019. Predicting Alzheimer's disease using LSTM. *IEEE Access* 7 (2019), 80893–80901.
[34] Heejung Jwa, Dongsuk Oh, Kinam Park, Jang Mook Kang, and Heuiseok Lim. 2019. exBAKE: Automatic fake news detection model based on bidirectional encoder representations from transformers (BERT). *Appl. Sci.* 9, 19 (2019), 4062.
[35] Abdullah Salem Khered, Ingy Yasser Hassan Abdou Abdelhalim, and Riza Theresa Batista-Navarro. 2022. Building an ensemble of transformer models for Arabic dialect classification and sentiment analysis. In *Proceedings of the 7th Arabic Natural Language Processing Workshop (WANLP'22)*. 479–484.
[36] Mingyang Li, Wanzhong Chen, and Tao Zhang. 2017. Automatic epileptic EEG detection using DT-CWT-based non-linear features. *Biomed. Sig. Process. Contr.* 34 (2017), 114–125.
[37] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
[38] S. Lock. 2022. What is AI chatbot phenomenon ChatGPT and could it replace humans? *'Book What is AI Chatbot Phenomenon ChatGPT and Could it Replace Humans* (2022). https://www.theguardian.com/technology/2022/dec/05/what-is-ai-chatbot-phenomenon-chatgpt-and-could-it-replace-humans. Accessed 09-07-2023.
[39] Antony Mayfield. 2008. What is social media. (2008). http://crmxchange.com/uploadedFiles/White_Papers/PDF/What_is_Social_Media_iCrossing_ebook.pdf.
[40] Muhammad Mujahid, Ernesto Lee, Furqan Rustam, Patrick Bernard Washington, Saleem Ullah, Aijaz Ahmad Reshi, and Imran Ashraf. 2021. Sentiment analysis and topic modeling on tweets about online education during COVID-19. *Appl. Sci.* 11, 18 (2021), 8438.
[41] Zineb Nassr, Nawal Sael, and Faouzia Benabbou. 2019. A comparative study of sentiment analysis approaches. In *Proceedings of the 4th International Conference on Smart City Applications*. 1–8.
[42] Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2013. How to construct deep recurrent neural networks. *arXiv preprint arXiv:1312.6026* (2013).
[43] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the Association for Computational Linguistics Meeting*. NIH Public Access, 2359.
[44] Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. 2023. On the effect of dropping layers of pre-trained transformer models. *Comput. Speech Lang.* 77 (2023), 101429.
[45] Vladimir Svetnik, Andy Liaw, Christopher Tong, J. Christopher Culberson, Robert P. Sheridan, and Bradley P. Feuston. 2003. Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* 43, 6 (2003), 1947–1958.
[46] Alper Kursat Uysal and Serkan Gunal. 2014. The impact of preprocessing on text classification. *Inf. Process. Manag.* 50, 1 (2014), 104–112.
[47] Eva A. M. van Dis, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L. Bockting. 2023. ChatGPT: Five priorities for research. *Nature* 614, 7947 (2023), 224–226.
[48] Karsten Wenzlaff and Sebastian Spaeth. 2022. *Smarter than Humans? Validating how OpenAI's ChatGPT Model Explains Crowdfunding, Alternative Finance and Community Finance.* Available at SSRN: https://ssrn.com/abstract=4302443; http://dx.doi.org/10.2139/ssrn.4302443
[49] Nurulhuda Zainuddin and Ali Selamat. 2014. Sentiment analysis using support vector machine. In *Proceedings of the International Conference on Computer, Communications, and Control Technology (I4CT'14)*. IEEE, 333–337.
[50] Wenying Zheng and Qiang Ye. 2009. Sentiment classification of Chinese traveler reviews by support vector machine algorithm. In *Proceedings of the 3rd International Symposium on Intelligent Information Technology Application*. IEEE, 335–338.
[51] Xujuan Zhou, Xiaohui Tao, Jianming Yong, and Zhenyu Yang. 2013. Sentiment analysis on tweets for social events. In *Proceedings of the IEEE 17th International Conference on Computer Supported Cooperative Work in Design (CSCWD'13)*. IEEE, 557–562.

Received 9 March 2023; revised 29 May 2023; accepted 12 June 2023