Journal of Engineering Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Engineering Research

journal homepage: www.journals.elsevier.com/journal-of-engineering-research

Unraveling user perceptions and biases: A comparative study of ML and DL models for exploring twitter sentiments towards ChatGPT

Mohammed Rashad Baker [a,*], Anil Utku [b]

[a] Software Department, College of Computer Science and Information Technology, University of Kirkuk, 36001 Kirkuk, Iraq
[b] Department of Computer Engineering, Faculty of Engineering, Munzur University, 62100 Tunceli, Turkey

ARTICLE INFO

Keywords:
Sentiment Analysis
ChatGPT
Deep Learning
CNN
Bi-LSTM

ABSTRACT

ChatGPT is a powerful chatbot that used to generate human-like text. Chat GPT, developed by OpenAI, can answer many questions honestly, like a personal teacher who knows almost everything. It can perform various tasks such as question and answer, solving mathematical equations, writing text, debugging, and translating between languages. Compared to traditional chatbots, ChatGPT has been adopted by many users since its introduction. Some users feel that ChatGPT will override most content creation professions. This study analyzed users' feelings about ChatGPT by analyzing tweets shared about ChatGPT. For this purpose, a hybrid Deep Learning (DL) model was developed using Convolutional Neural Networks (CNN) and Bi- Long Short-Term Memory (Bi-LSTM) models. The study has been compared with a number of DL and Machine Learning algorithms, LSTM, Bi-LSTM, CNN, Gated Recurrent Unit, Random Forests and Support Vector Machines. The experimental outcomes demonstrated that, the FastText-trained CNN-Bi-LSTM model exceeded the other models in terms of accuracy, reaching 96.59%.

1. Introduction

The content of user-generated has been increased at an peerless rate due to the growth of social media platforms, most especially Twitter [1]. The substantial increase in data presents a noteworthy chance for scholars and analysts to acquire a profound understanding of public sentiments, opinions, and dominant trends [2]. Sentiment analysis, developed as a powerful tool, employs techniques to extract information from social media data, distinguishing common sentiments on specific topics or events and providing valuable awareness into public sentiment [3,4].

The capturing ability of the complications of human language and generate comprehensible responses verified by natural language processing (NLP) model developed by OpenAI[1] called ChatGPT [5]. However to fully utilize the capabilities of ChatGPT for pragmatic implementations, it is critical to assess its efficacy in authentic contexts, such as the dynamic environment of Twitter, where users carry an general spectrum of sentiments and viewpoints by entirely utilizing the abilities of ChatGPT for practical applications [6]. ChatGPT differentiates itself as a valuable tool across many domains—including information retrieval, content generation, and customer support—utilizing its exceptional ability for smoothing cohesive and captivating conversation [7]. As what can be happen with any Artificial Intelligence (AI) technology, it also causes individuals and society to express apprehensions and queries. It is imperative to understand the sentiments associated with ChatGPT and similar AI technologies to assess public opinion and devise strategies to overcome potential complications [8].

The rise of advanced generative AI systems like ChatGPT has sparked both exciting and apprehension. Alternatively, promoters for these technologies show the capacity to fundamentally transform numerous sectors and elevate the quality of user experiences. Though, ethical insinuations, privacy concerns, and the effects of AI on human work are also subjects of trepidation. These concerns frequently create from fears regarding AI systems' potential replacement of human work, the possibility of AI producing biased or detrimental outputs, and its capability to rigged or cheat users [9].

This study loads great importance for a number of reasons. To begin with, it facilitates the comprehension of public sentiment and opinion, given that Twitter is a vastly utilized platform where users often articulate their viewpoints. Examining these sentiments may provide significant insights into how the general public views ChatGPT. Secondly, analysis of the sentiments and feedback stated on Twitter can enable the

* Corresponding author.
E-mail address: mohammed.rashad@uokirkuk.edu.iq (M.R. Baker).
[1] https://openai.com/

https://doi.org/10.1016/j.jer.2023.11.023
Received 13 August 2023; Received in revised form 11 November 2023; Accepted 23 November 2023
Available online 28 November 2023

identification of potential areas for enhancement in ChatGPT. This can provide guidance for subsequent development endeavors aimed at improving the efficacy of the system and ensuring user comfortably. Thirdly, this study contributes to the field of ML and DL by implementing and assessing various models in a real-world context. The findings can notify the design of more actual models for sentiment analysis. Lastly, considerate public sentiment towards ChatGPT can inform actual approaches for its placement and implementation. Positive sentiments can be leveraged for promotional purposes, while handling negative sentiments can help ease potential reactions. Therefore, this study holds significance not only for the development and improvement of ChatGPT but also for advancing Machine Learning (ML)/Deep Learning (DL) techniques and informing actual deployment policies.

Considerate the sentiment scenery surrounding ChatGPT is essential for developers, policymakers, and researchers. It can inform approaches for optimizing AI system design, addressing ethical concerns, and building trust with users. Furthermore, the analysis presented herein contributes to the broader discourse on responsible AI development and placement, with the possible to inform the creation of principles and guidelines that indorse the ethical and advantageous take advantage of AI technologies.

In this study, the experiments explore the prediction and sentiment analysis of ChatGPT on Twitter, a dominant social media platform where users openly carry their opinions, concerns, and experiences. The aim is to collect insights from the sentiments expressed by users, thus providing a deeper thoughtful of the public perception of ChatGPT and the challenges it challenges acceptance and ethical considerations.

The main purpose of this study is to analyze users' sentiments towards ChatGPT over the inspection of tweets specifically related to ChatGPT. A hybrid DL model is being developed by applying both CNN and Bi-LSTM architectures. A comprehensive comparison will be conducted between the developed model and prominent models such as Support Vector Machines (SVM), Random Forests (RF), Gated Recurrent Unit (GRU), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Bidirectional LSTM (Bi-LSTM). Furthermore, experimental studies will be carried out employing FastText, Global Vectors for Word Representation (GloVe), and Term Frequency - Inverse Document Frequency (TF-IDF) techniques.

The main contributions of this study content:

- Suggesting a novel dataset of sentiment-labeled tweets about ChatGPT, which may verify useful for upcoming social media sentiment analysis research.

- Showing a comprehensive comparison of many ML and DL models for sentiment analysis, describing their individual strengths and weaknesses.

- Investigating the results of these models and discussing the factors that impact the sentiments of Twitter users about ChatGPT.

The remainder of this study is separated into the following sections. A review of the current state of research on sentiment analysis and conversational agents is given in Section 2. Section 3 outlines the procedure for gathering and annotating the dataset. Section 4 details the methodology and experimental configuration, giving and analyzing the gained results. Finally, in this study's concluding section, Section 5, important discoveries are highlighted and possible directions for future research are recommended.

2. Related works

The inspection of sentiments carried by users on social media platforms, including Twitter, for emerging technologies in the present day has garnered considerable scholarly attention [10]. Scholars have explored the application of ML models in predicting sentiments and opinions. In addition, the comprehension of public opinion about AI language models such as ChatGPT has acquired great importance. This section reviews the relevant research, specifically highlighting sentiment analysis within the context of developing technologies and the use of ML models. Besides, it investigates into the sentiments used by different people regarding ChatGPT.

Sentiment analysis extracts and evaluates opinions from textual data to find the sentiment or emotional tone expressed by individuals. Various research studies have used sentiment analysis methods to study how people feel about new technology on social media networks like Twitter.

In recently, significant emphasis has been placed on the sentiment related to AI language models, especially ChatGPT, which are known for their conversational capabilities and human-like responses. Researchers explored the perspectives of individuals regarding the features and effects of ChatGPT. ChatGPT received considerable interest mainly because of its exceptional versatility and wide array of practical applications.

ChatGPT is a sentiment analysis instrument that has been developed with the explicit purpose of mitigating substantial cybersecurity concerns, according to Okey et al. [11] Kasneci et al. [12], discussed the benefits and challenges of using language models in educational applications from both educator and student perspectives.

Iftikhar et al. [13] mentioned that ChatGPT demonstrated the ability to provide medical information. Mijwil et al. [14] raised concerns that the use of ChatGPT in medical applications may create concerns about data privacy. AI methods are used to determine users' feelings towards ChatGPT on social media. By using ML and DL methods, a general perspective on the public's opinions on certain issues is obtained through social shares.

Hasan et al. [15] presented a hybrid method using SentiWordNet, W-WSD and TextBlob methods for sentiment analysis. The developed method was compared with Naïve Bayes (NB) and SVM. Experimental results showed that W-WSD was more successful than other compared methods. The study confirmed the efficiency of their hybrid method in capturing sentiments in political texts.

Neethu et al. [16] analyzed the sentiments in Twitter posts associated to electronic products using an ML approach. Domain-specific characteristics should be combined into sentiment classification, the authors highlighted. They put forth an unique feature vector to differentiate positive or negative tweets and gather individuals' opinionsregarding products. Through the application of ML, their researchassisted in investigating emotion in the context of electronic products.

Le and Nguyen [17] suggested a methodology for an analysing sentiment on social with the primary importance of their research was feature selection via techniques including Bigram extraction and Information Gain. Furthermore, they proposed a model for sentiment analysis that utilized NB and SVM. The authors highlighted the model's worthand precision in sentiment analysis, specifically about social media data.

Additionally, Jain et al. [18] introduced an overall and sequential description of sentiment analysis using ML. they proposed a framework for text analysis, leading to an enhanced, more flexible, and scalable analysis technique by utilizing Apache Spark. They implemented NB and DT algorithms by introducing a comprehensive methodology for analyzing Twitter data sentiment.

For classifying Twitter data, Naresh and Krishna [19] proposed an optimization-based ML algorithm. Their strategy included the acquisition and preprocessing of data, the extraction of features, and the classification of data utilizing a variety of ML algorithms. They found that sequential minimal optimization with decision trees, achieved high accuracy compared to other algorithms by highlighting the importance of optimization techniques in sentiment analysis.

Sentiment analysis classifications on movie review data were performed by Rahman et al. [20] through the application of several ML techniques. Classifiers such as Bernoulli NB, DT, SVM, maximum entropy, and multinomial NB were compared, and it was demonstrated that the Multinomial NB classifier defeated others in terms of overall performance.

A common discussions have been thought about people's sentiments and opinions about ChatGPT. While AI language models are observed by researchers as powerful tools that increase productivity and communication, concerns about their ethical implications, potential biases, and impact on human work are stated by others. Valued visions into the public awareness of AI language models and their general role are presented by these sentiments.

A practical method for indicating sentiment analysis on tick-borne disease material using Natural Language Processing (NLP) techniques was provided by Susnjak [21]. A pipeline involving of data collection, preprocessing, annotation, feature extraction, model selection, evaluation, and deployment was suggested. Some of the challenges and best sentiment analysis applications in this domain were also discussed.

In addition to traditional NLP techniques, growing attention has been given to using large language models (LLMs) like ChatGPT for Sentiment Analysis. ChatGPT's understanding capabilities were studied by Zhong et al. [22] through testing it against the most notable GLUE benchmark and compared it to four sample fine-tuned BERT-style models. It was found that ChatGPT achieved competitive results on some tasks, such as sentiment analysis and natural language inference, but lagged behind on others, such as question answering and textual implication. The reason was qualified to differences in pre-training objectives and data sources between ChatGPT and BERT-style models.

Leiter et al. [23] conducted sentiment analysis about ChatGPT using datasets consisting of 330k tweets and 150 academic articles obtained from arxiv and Google Scholar.

Kocoń et al. [24], evaluated the effectiveness of ChatGPT in terms of sentiment analysis, emotion recognition, aggression, and posture detection. Experimental results showed that ChatGPT failed to detect hate speech and fake news, but succeeded in detecting sarcasm and irony. Wang et al. [25], conducted a study on ChatGPT's ability to detect emotions conveyed in text using IMDb and Amazon datasets. The results of ChatGPT were compared with TextBlob and VADER. Experimental results showed that ChatGPT was not good enough at detecting emotions.

Xie et al. [26] conducted a study on the use of ChatGPT in stock forecasting. Historical stock price data and tweets were used as distinct datasets, and the functionalities of ChatGPT within this actual field were investigated and assessed using a zero-shot methodology. Sentiment scores for each tweet using ChatGPT and then incorporated those scores into a linear regression model to forecast the movement of the stock price. According to their findings, ChatGPT outperformed state-of-the-art methods that rely solely on visual or textual characteristics or combine the two.

Zhu et al. [27]. investigated the ability of ChatGPT to regenerate label annotations generated by humans in social computing tasks. Humor detection (Haha), sarcasm detection (SARC), emotion recognition (EmoBank), and sentiment analysis (SST-2) were used as four different datasets. Predictions of ChatGPT were compared to those of baseline models, including BERT and RoBERTa. On the majority of datasets, ChatGPT outperformed the baselines, except for humor detection.

Joshi et al. [28] used a quantitative methodology to display the significant unreliability of ChatGPT when it comes to answering a wide array of questions related to various subjects within undergraduate computer science. A dataset comprising 1,000 inquiries spanning several subjects, including algorithms, data structures, programming languages, databases, operating systems, and networks, was used. ChatGPT's answers were compared with correct answers and with baseline models such as GPT-2 and GPT-3. It was stated that ChatGPT had a low accuracy rate compared to GPT-2 and GPT-3.

A complete examination of the abilities of LLMs for a variety of sentiment analysis tasks was shown by Zhang et al. [29] tasks comprised traditional sentiment classification, Aspect-Based Sentiment Analysis (ABSA), and multifaceted analysis of subjective texts. Sentiment classification (IMDb), ABSA, Opinion Role Labeling (ORL), Opinion Target Extraction (OTE), and Opinion Relation Extraction (ORE) were used as five different datasets. ChatGPT's predictions were compared with human annotations and with baseline models such as BERT and XLNet. It was quantified that ChatGPT attained notable results on sentiment classification and ABSA but struggled with more fine-grained tasks such as opinion role labeling, opinion target extraction, and opinion relation extraction.

Lastly, Golubev et al. [30] A corpus of 10,000 news articles annotated with sentiments to entities mentioned in the texts was collected. ChatGPT was evaluated alongside other models such as BERT-RuSentRel, RuBERT, and SentiRuEval-2016 on entity-level sentiment analysis and relation-level sentiment analysis as two different tasks.

In contrast to the work by Haque et al. [31], our research endeavors to explore and evaluate user sentiments surrounding ChatGPT through a distinctive approach. While they focus on a mixed-method study utilizing a substantial dataset of tweets from early ChatGPT users and employing topic modeling coupled with qualitative sentiment analysis, our investigation diverges in methodology and objectives.

Sentiment analysis of people's opinions on social media platforms, particularly Twitter, regarding new technologies has attracted significant research attention. ML models, including supervised learning algorithms and DL architectures, have been employed to predict sentiments and opinions accurately. Furthermore, conducting sentiment analysis on individuals' comments concerning AI language models such as ChatGPT provides significant insights into the general public's attitude and concerns towards these technologies.

3. Methodology

Here, the methodology employed in this article is presented by comparing and analyzing various models for sentiment. The fundamental purpose of this study is to evaluate the performance of various AI and DL models, in addition to word embedding methods, when analyzing the sentiment conveyed in textual data. This section thoroughly describes the models and methodologies assessed in the following examination.

To accomplish the research aim, a comparative examination of various sentiment analysis models will be undertaken. The models' performance will be assessed according to their precision in classifying and identifying the sentiment expressed in textual data. By conducting this analysis, one may identify the merits and demerits of each model and subsequently generate well-informed suggestions concerning their suitability for sentiment analysis tasks.

3.1. Dataset

The research employed a dataset comprising 217,623 tweets compared to the initial 219,294 tweets about ChatGPT, after which duplicate entries were removed. Among these, there are 106,695 negative tweets, 55,754 positive tweets, and 55,174 neutral tweets. You can find a sample from the dataset in Table 1.

The word cloud of the dataset is shown in Fig. 1. 2. The dataset must be converted into a natural language processing process in the data preprocessing stage. For this purpose, hashtags, usernames, punctuations, emojis, special characters and links in tweets have been removed. Uppercase letters have been converted to lowercase letters. Then tokenization, stopwords removal, stemming and lemmatization were applied. Data were standardized, tokenized and vectorized using TextVectorization. Standardization refers to removing punctuation and HTML codes to simplify the dataset. Tokenization refers to the splitting of tweets into tokens. With tokenization, sentences are broken down into words. Vectorization enables words, sentences or documents to be expressed as vectors. TfidfVectorizer was used to convert tweet texts to numerical features.

Stemming is the process of removing suffixes from a word. Stemming

and lemmatization are used to access the root of the word. In both operations, it aims to purify each word from its suffixes and transform it into a common base or root. Stemming algorithms endeavor to reduce a word by considering prevalent prefixes and suffixes, potentially found in an affixed word, either at the word's start or end. Porter Stammer algorithm was used for stemming. Lemmatization is the morphological rooting of the word. WordNetLemmatizer was used for lemmatization. The class labels of the tweets were encoded using LabelEncoder to be Class 0 for negative, Class 1 for positive, and Class 2 for neutral.

After data pre-processing, training and test sets were split (80%:20%). 10% of the training dataset was split as validation data to optimize the model parameters. Grid search was used for hyperparameter tuning. The test dataset includes 6002 negative, 5476 positive and 5287 neutral tweets.

3.2. ML models

Many ML models like RF were used to perform sentiment analysis. It is an ensemble learning algorithm that builds multiple decision trees and combines their predictions to arrive at ultimate classifications. The RF model utilizes a collection of decision trees to analyze and classify textual data based on the sentiment effectively expressed [32].

In addition, SVM model stands as one of most used models for NLP tasks. It is a robust supervised learning algorithm extensively utilized in classification tasks, notably in applications such as sentiment analysis. SVM's primary objective is to identify an optimal hyperplane, which maximizes the distinction between various classes by transforming data into a high-dimensional feature space. In the context of sentiment analysis, SVM models demonstrate effectiveness in classifying text according to the conveyed sentiment [33].

3.3. DL models

Many DL models like CNN are used in sentiment analysis classifications. CNN is a deep learning architecture known for its remarkable performance in various natural language processing applications, among them sentiment analysis. CNNs achieve precise sentiment classification by utilizing convolutional layers to extract local features from text and capture relevant patterns [34].

Another model is LSTM, it is holds the ability to identify long-range dependences within sequential data, making them compatible for sentiment analysis takes that require conception of the sequential dynamics and context of sentiment expression. It also holds the ability to discriminate and retain patterns over a period of time, making them highly suitable for sentiment analysis in text [4].

Another model is Bi-LSTM, which is an expansion of the LSTM architecture, enables the processing of the input sequence in both forward and reverse directions. equential dynamics and contextual subtleties inherent to textual data can be captured more efficiently using bidirectional processing in Bi-LSTM. Bi-LSTM models are capable of accurately discerning sentiment in both past and future contexts by recognizing the complex relationships and interdependencies present in the text. Bi-LSTM is exceptionally well-suited for sentiment analysis tasks that necessitate an exhaustive understanding of sentiment expressions within sequential data [35].

Table 1
A sample from the dataset.

Tweet	Class
ChatGPT: Optimizing Language Models for Dialog.	Neutral
Try talking with ChatGPT, our new AI system wh.	Positive
ChatGPT: Optimizing Language Models for Dialog.	Neutral
THRILLED to share that ChatGPT, our new model.	Positive
As of 2 min ago, @OpenAI released their ne.	Negative

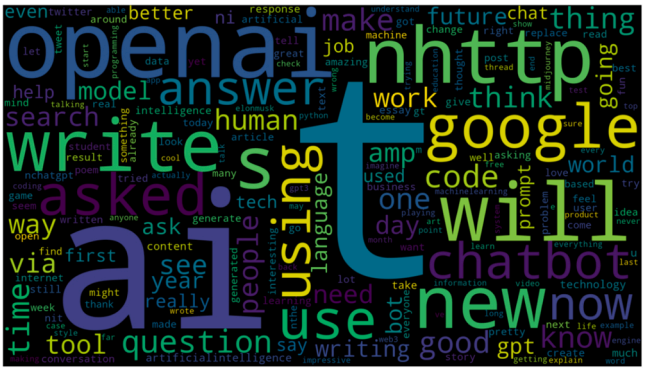

Fig. 1. Word cloud of the dataset.

3.4. Word embedding techniques

There were many word embedding techniques, one of these techniques is FastText. FastText, uses continuous vectors to represent words in a high-dimensional space. The system leverages sub-word information to process words that are not in its vocabulary effectively and to capture subtle morphological details. FastText has been extensively implemented in sentiment analysis tasks to improve classification performance and word representation [36].

Another technique is the Global Vectors for Word Representation (GloVe). GloVe is a well-known word embedding method implemented in sentiment analysis. It leverages co-occurrence statistics to generate word vectors that encapsulate semantic relationships. By capturing the global context of words, GloVe enhances the sentiment analysis process and contributes to accurate classification [37].

Term Frequency - Inverse Document Frequency (TF-IDF), is another word embedding technique that is widely adopted weighting scheme in text mining and information retrieval, is prominently applied in various tasks, including sentiment analysis. It assesses the significance of a term within a document by accounting for its frequency in that document and its scarcity across the entire corpus. TF-IDF adeptly captures the distinguishing influence of words and contributes to the process of sentiment classification [37].

3.5. Developed CNN-Bi-LSTM model

This study aimed to analyze the users' sentiments about ChatGPT by analyzing the tweets shared by users about ChatGPT. To achieve that, a hybrid DL model comprising CNN and Bi-LSTM models was constructed. A hybrid model, CNN-Bi-LSTM, integrates the respective advantages of CNN and Bi-LSTM. By integrating convolutional and recurrent layers, CNN-Bi-LSTM models can efficiently capture long-term dependencies and contextual information in text, thereby augmenting the accuracy of sentiment analysis [37].

The main goal of developing the model is to improve the performance of both Bi-LSTM and CNN models. CNN is renowned for its effectiveness in feature extraction, while Bi-LSTM facilitates bidirectional data processing. This means that input data is initially processed in its original sequence and then in reverse during the second stage. The aim was to enhance classification accuracy by combining CNN and Bi-LSTM models in a hybrid approach. Figure 6 illustrates the architecture of the resulting CNN-Bi-LSTM model.

The developed model performs data pre-processing on the raw Tweet dataset. Then, important information is extracted from the tweet texts using the FastText. The tweet dataset is presented to CNN as input after the pre-processing stages in the proposed model. CNN consists of Conv1D layers with filter size 32, kernel size 3, padding same, and ReLU activation function parameters. In the fully connected layer, the data passing through the convolutional layer and the pooling layer and in the form of a matrix is turned into a flat vector. Bi-LSTM layers consist of 32

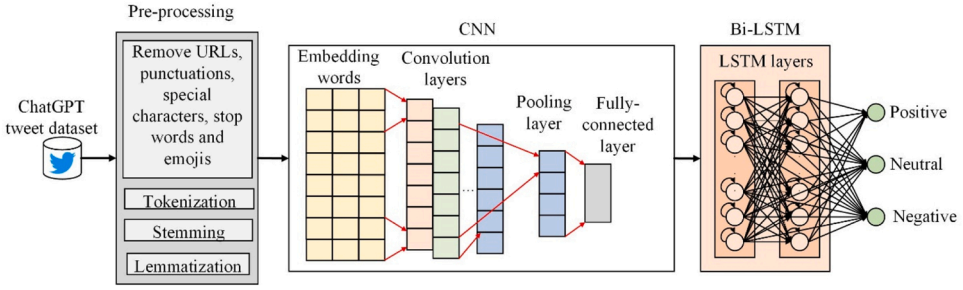

Fig. 2. The structure of the developed hybrid model.

LSTM neurons. Softmax was used as the activation function in the Dense layer. Hyperparameter analysis was conducted on all the employed models through GridSearch. This process involved evaluating each model with the optimal set of parameters, ensuring the best possible configurations were used.

GridSearch is a common hyperparameter optimization method used to determine the hyperparameters of ML and DL models. Essentially, it creates a grid to try different combinations of hyperparameters to optimize the performance of a particular algorithm. The GridSearch method takes a list or range of specific hyperparameters and creates different combinations of these parameters to evaluate the model's performance. It then trains the model using these combinations and evaluates its performance with a metric (e.g., accuracy, F1 score, RMSE). It selects the hyperparameter combination that gives the best performance among the results. The hyperparameters of the developed model are shown in Table 2.

Table 2
Hyperparameters of the developed model.

Parameters	Value
Epoch	20
Learning_rate	0.1
Filters in Conv1D layer	32
Kernel_size in Conv1D layer	3
Padding in Conv1D layer	Same
Activation function in Conv1D layer	Relu
Pool_size in MaxPooling1D layer	2
Number of neurons in the LSTM layer	32
Dropout value	0.4
Number of neurons in the dense layer	3
Activation function in the Dense layer	Softmax

4. The experimental results

In this study, a hybrid DL model was developed for sentiment analysis of tweets shared about ChatGPT. The developed CNN-Bi-LSTM hybrid model was compared with RF, SVM, CNN, LSTM and Bi-LSTM. In experimental studies, FastText, GloVe and TF-IDF techniques were applied. FastText had more successful results than other techniques. Continuation of this section, confusion matrices according to FastText word embedding technique are presented for all applied models.

In experimental studies, 10-fold cross validation was used. With cross validation, the data set was randomly divided into 10 clusters and experimental studies were conducted for each cluster. Results were calculated by averaging all experiments. Experimental results were obtained for each model based on accuracy, precision, recall, and F-score metrics. The weighted average was calculated for each model's precision, recall and F-score.

The confusion matrix of RF is shown in Table 3. The table shows that RF accurately classified 14,760 tweets out of 16,765, totaling 0.8804%. This underscores the model's robust performance in accurately classifying sentiments from tweets, a critical element in sentiment analysis undertakings.

Table 4 shows SVM model confusion matrix. The table indicates that SVM classified 14,776 tweets out of 16,765 with an accuracy rate of 0.8813. This shows the ability of SVM, which is crucial forsentiment analysis applications, to precisely label tweets with sentimentlabels.

Table 5 illiterates CNN confusion matrix. The table presents that CNN classified 14,852 tweets out of 16,765 accurately, for an imposing accuracy rate of 0.8858. The notable degree of precision exhibited by CNN in classifying tweets consistent with their distinct sentiment categories highlights the network's efficacy in sentiment analysis classification.

Table 6 shows the confusion matrix of GRU. GRU correctly classified 15,222 out of 16,765 tweets, achieving an accuracy of 0.9079. The accuracy rate obtained showed that GRU was successful in distributing sentiment labels to tweets.

Table 7 shows the confusion matrix of LSTM. LSTM correctly classified 15,303 out of 16,765 tweets, achieving an accuracy of 0.9127. The accuracy rate obtained showed that LSTM was successful in distributing sentiment labels to tweets.

Table 8 shows the confusion matrix of Bi-LSTM. Bi-LSTM correctly classified 15,532 out of 16,765 tweets, achieving an accuracy of 0.9264. The accuracy rate obtained showed that LSTM was successful in distributing sentiment labels to tweets.

Table 9 shows the confusion matrix of CNN-Bi-LSTM. CNN-Bi-LSTM correctly classified 16,203 out of 16,765 tweets, achieving an accuracy of 0.9664. Experimental results showed that CNN-Bi-LSTM has higher accuracy than compared models.

Table 10 shows comparative experimental results for accuracy, precision, recall, and F-score. The table shows that the developed CNN–Bi-LSTM hybrid model trained with FastText has a higher classification performance than the compared models. The CNN–Bi-LSTM model had the best performance compared to other models in terms of all compared metrics and word embedding techniques. After the developed model, Bi-LSTM, LSTM, GRU, CNN, SVM, and RF were successful, respectively.

Figs. 3, 4 and 5 show the accuracy and loss graphs for the developed model trained using FastText, GloVe, and TF-IDF.

Fig. 3 shows the situation where the CNN-Bi-LSTM model is trained with FastText. As seen in the graph, the model's accuracy performance improves impressively during training. Although the accuracy rate is low starting at the first epoch, it increases rapidly over time, reaching approximately 0.9664 by the 10th epoch. These results show that the model can make highly accurate predictions on the training data. Additionally, when the loss graph is examined, the loss values of the

Table 3
RF's confusion matrix.

		Actual		
		Negative	Positive	Neutral
Predicted	Negative	5548	83	408
	Positive	59	4760	427
	Neutral	395	633	4452

Table 4
SVM's confusion matrix.

		Actual		
		Negative	Positive	Neutral
Predicted	Negative	5484	98	387
	Positive	79	4885	493
	Neutral	439	493	4407

Table 5
CNN's confusion matrix.

		Actual		
		Negative	Positive	Neutral
Predicted	Negative	5498	64	390
	Positive	112	4865	408
	Neutral	392	547	4489

Table 6
GRU's confusion matrix.

		Actual		
		Negative	Positive	Neutral
Predicted	Negative	5638	54	360
	Positive	63	4985	328
	Neutral	301	437	4599

Table 7
LSTM's confusion matrix.

		Actual		
		Negative	Positive	Neutral
Predicted	Negative	5694	51	352
	Positive	63	4994	320
	Neutral	245	431	4615

model continue to decrease as the training progresses. This decrease indicates that the model fits the training data better and makes its predictions more accurate.

Fig. 4 shows the situation where the CNN-Bi-LSTM model is trained with GloVe. In this graph, the accuracy performance of the model is quite impressive. Although it starts off low when training begins, it improves over time and achieves an accuracy of around 0.9654 by the 10th epoch. The loss chart shows that initially, high loss values are decreasing. Predictions are improved as the model matches more closely with the training data.

The scenario in which the CNN-Bi-LSTM model is trained using TF-IDF is illustrated in Fig. 5. Another similar trend is evident in this graph. The model's accuracy performance exhibits an upward enhancement throughout the training process, resulting in an approximate value of 0.9644 at the 10th epoch. The loss chart illustrates how the model decreases the initially significant loss values.

The figures above show that the model successfully fits the training data in all three cases, and its accuracy performance improves. Furthermore, the loss graph illustrates that the model improves in learning as the training process advances, resulting in better predictions. The outcomes demonstrate the effectiveness of the CNN-Bi-LSTM model across various embedding techniques (FastText, GloVe, TF-IDF) during training.

5. Conclusion and future works

In this study, the hybrid CNN-Bi-LSTM model outperforms all other models by attaining the highest accuracy, precision, recall, and F-score values. The achievement can be assigned to the novel combination of CNN and Bi-LSTM networks, which efficiently played on each of the merits of both algorithms. The CNN demonstrates superior skills in capturing local features, whereas the Bi-LSTM excels at modeling sequential data. In summary, the developed hybrid model has been successfully applied to identify complex features in tweets.

Table 8
Bi-LSTM's confusion matrix.

		Actual		
		Negative	Positive	Neutral
Predicted	Negative	5739	37	327
	Positive	49	5107	274
	Neutral	214	332	4686

Table 9
CNN-Bi-LSTM's confusion matrix.

		Actual		
		Negative	Positive	Neutral
Predicted	Negative	5891	19	174
	Positive	22	5296	97
	Neutral	89	161	5016

Table 10
Comparative experimental results.

Word Embedding Technique	Model	Accuracy	Precision	Recall	F-Score
FastText	RF	0.8804	0.8794	0.8785	0.8789
	SVM	0.8813	0.8797	0.8797	0.8797
	CNN	0.8858	0.8847	0.8844	0.8845
	GRU	0.9079	0.9068	0.9064	0.9065
	LSTM	0.9127	0.9116	0.9111	0.9113
	Bi-LSTM	0.9264	0.9254	0.9250	0.9251
	CNN-Bi-LSTM	**0.9664**	**0.9662**	**0.9657**	**0.9659**
GloVe	RF	0.8789	0.8780	0.8780	0.8780
	SVM	0.8799	0.8783	0.8783	0.8783
	CNN	0.8844	0.8832	0.8830	0.8830
	GRU	0.9065	0.9053	0.9050	0.9051
	LSTM	0.9113	0.9101	0.9097	0.9098
	Bi-LSTM	0.9250	0.9241	0.9236	0.9238
	CNN-Bi-LSTM	**0.9654**	**0.9651**	**0.9647**	**0.9648**
TF-IDF	RF	0.8782	0.8773	0.8763	0.8767
	SVM	0.8792	0.8775	0.8775	0.8775
	CNN	0.8837	0.8825	0.8823	0.8823
	GRU	0.9058	0.9046	0.9043	0.9044
	LSTM	0.9106	0.9094	0.9089	0.9091
	Bi-LSTM	0.9243	0.9233	0.9229	0.9230
	CNN-Bi-LSTM	**0.9648**	**0.9648**	**0.9641**	**0.9644**

In the developed hybrid model, different word embedding methods such as FastText, GloVe and TF-IDF were used. In this way, the model's capacity to analyze complex emotions in tweets was increased.

Experimental results showed that recurrent neural network models such as Bi-LSTM, GRU and LSTM are more successful than CNN, RF and SVM. Traditional machine learning methods such as RF and SVM were not successful enough due to the high size and complex structure of tweets.

The real-world application of this study are notable. With a highly accurate sentiment analysis model in place, researchers and practitioners can gain valuable insights into public sentiment regarding ChatGPT and similar AI-supported technologies. These awarenesses are crucial for understanding how such technologies are perceived and integrated into daily life and various business applications. In essence, this study highlights the strengths and weaknesses of various models, aiding in the informed selection of the most suitable approach for sentiment analysis in social media contexts.

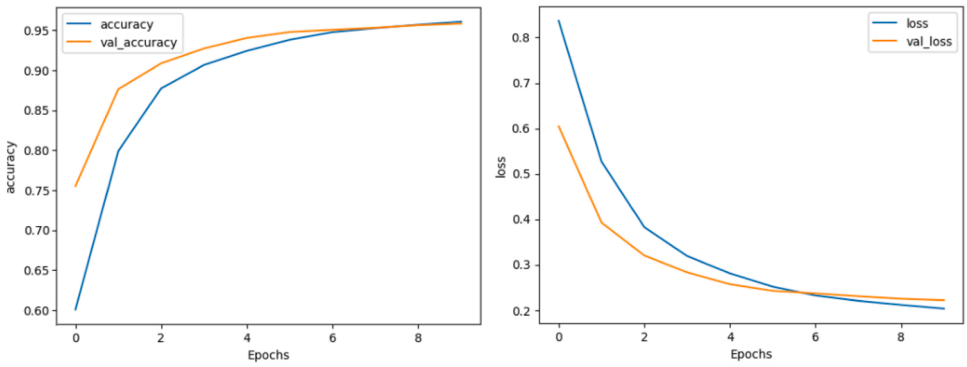

Fig. 3. Accuracy and loss graphs of developed CNN-BiLSTM model trained using FastText.

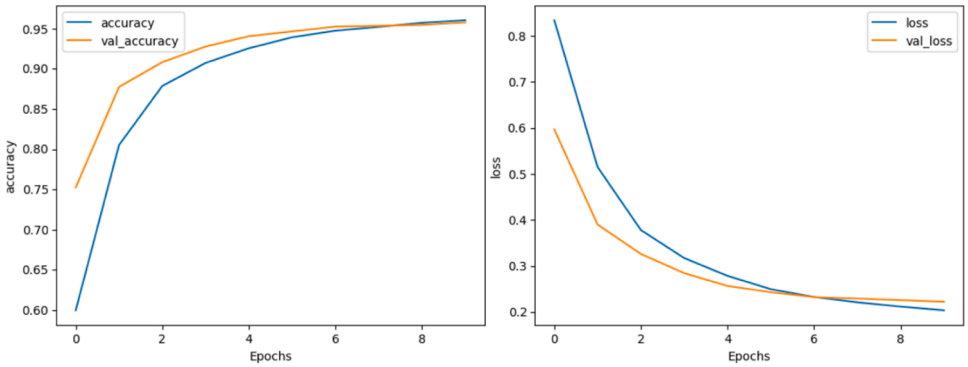

Fig. 4. Accuracy and loss graphs of developed CNN-BiLSTM model trained using GloVe.

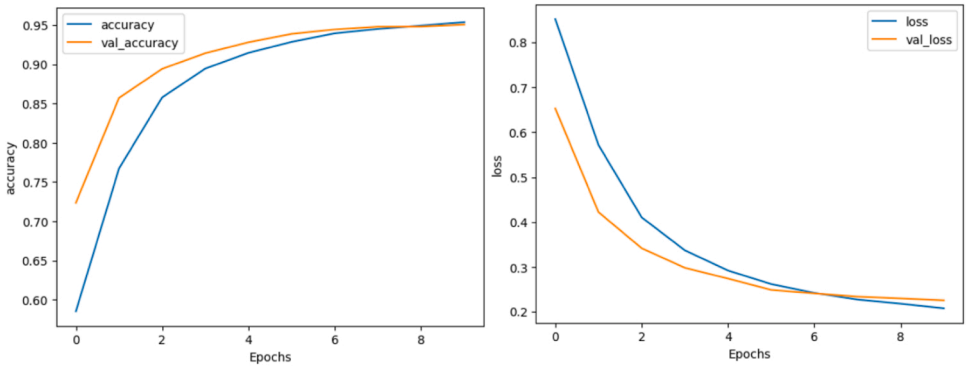

Fig. 5. Accuracy and loss graphs of developed CNN-BiLSTM model trained using TF-IDF.

Beyond the current findings, future research should focus on advanced neural networks, transferability to other domains, explainable AI, real-time analysis, and ethical considerations. Additionally, studying human-AI collaboration and long-term sentiment trends is essential to keep sentiment analysis relevant and adaptable in the evolving social media and AI landscape.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] W. Duan, Q. Cao, Y. Yu, S. Levy, Mining online user-generated content: Using sentiment analysis technique to study hotel service quality, Proc. Annu. Hawaii Int. Conf. Syst. Sci. (2013) 3119–3128, https://doi.org/10.1109/HICSS.2013.400.

[2] M.R. Baker, K. Jihad, Y. Taher, Prediction of People Sentiments on Twitter using Machine Learning Classifiers During Russian Aggression in Ukraine, Jordan. J. Comput. Inf. Technol. (2023) 1, https://doi.org/10.5455/jjcit.71-1676205770.

[3] T. Hu, et al., Revealing public opinion towards covid-19 vaccines with twitter data in the united states: Spatiotemporal perspective, J. Med. Internet Res. vol. 23 (9) (2021), e30854, https://doi.org/10.2196/30854.

[4] M.R. Baker, E.Z. Mohammed, K.H. Jihad, Prediction of Colon Cancer Related Tweets Using Deep Learning Models. Intelligent Systems Design and Applications. ISDA 2022. Lecture Notes in Networks and Systems, Springer, Cham, 2023, pp. 522–532, https://doi.org/10.1007/978-3-031-27440-4_50.

[5] A. Borji, A Categorical Archive of ChatGPT Failures," arXiv Prepr. arXiv2302.03494, Feb. 2023, Accessed: Jun. 11, 2023. [Online]. Available: http://arxiv.org/abs/2302.03494.

[6] M. Heumann, T. Kraschewski, and M.H. Breitner, ChatGPT and GPTZero in Research and Social Media: A Sentiment-and Topic-based Analysis," SSRN, p.

4467646, 2023, Accessed: Jun. 11, 2023. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4467646.

[7] P.P. Ray, ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope, Internet Things Cyber-Phys. Syst. vol. 3 (2023) 121–154, https://doi.org/10.1016/j.iotcps.2023.04.003.

[8] Y.K. Dwivedi, et al., So what if ChatGPT wrote it?' Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy, Int. J. Inf. Manag. vol. 71 (2023), https://doi.org/10.1016/j.ijinfomgt.2023.102642.

[9] M. Sallam, ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns, Healthc. (Basel, Switz.) vol. 11 (6) (2023), https://doi.org/10.3390/healthcare11060887.

[10] K.H. Jihad, M.R. Baker, M. Farhat, M. Frikha, Machine Learning-Based Social Media Text Analysis: Impact of the Rising Fuel Prices on Electric Vehicles. Hybrid Intelligent Systems. HIS 2022. Lecture Notes in Networks and Systems, Springer, Cham, 2023, pp. 625–635, https://doi.org/10.1007/978-3-031-27409-1_57.

[11] O.D. Okey, E.U. Udo, R.L. Rosa, D.Z. Rodríguez, J.H. Kleinschmidt, Investigating ChatGPT and cybersecurity: a perspective on topic modeling and sentiment analysis, Comput. Secur. vol. 135 (. 2023), 103476, https://doi.org/10.1016/j.cose.2023.103476.

[12] E. Kasneci, et al., ChatGPT for good? On opportunities and challenges of large language models for education, Learn. Individ. Differ. vol. 103 (. 2023), 102274, https://doi.org/10.1016/j.lindif.2023.102274.

[13] L. Iftikhar, DocGPT: Impact of ChatGPT-3 on Health Services as a Virtual Doctor," *EC Pae*diatr., vol. 3, pp. 45–55, 2023, (Accessed: 26 October 2023). [Online]. Available: https://www.researchgate.net/profile/Muhammad-Iftikhar/publication/369013064_DocGPT_Impact_of_ChatGPT-3_on_Health_Services_as_a_Virtual_Doctor/links/6404151eb1704f343fa1c964/DocGPT-Impact-of-ChatGPT-3-on-Health-Services-as-a-Virtual-Doctor.pdf.

[14] M. Mijwil, M. Aljanabi, A.H. Ali, ChatGPT: exploring the role of cybersecurity in the protection of medical information, Feb. 2023, Mesop. J. Cyber Secur. vol (2023) 18–21, https://doi.org/10.58496/mjcs/2023/004.

[15] A. Hasan, S. Moin, A. Karim, S. Shamshirband, Machine learning-based sentiment analysis for twitter accounts, Math. Comput. Appl. vol. 23 (1) (. 2018) 11, https://doi.org/10.3390/mca23010011.

[16] M.S. Neethu, R. Rajasree, Sentiment analysis in twitter using machine learning techniques, *ICCCNT 2013*, 2013 4th Int. Conf. Comput., Commun. Netw. Technol. (2013), https://doi.org/10.1109/ICCCNT.2013.6726818.

[17] B. Le, H. Nguyen, Twitter sentiment analysis using machine learning techniques, Adv. Comput. Methods Knowl. Eng.: Proc. 3rd Int. Conf. Comput. Sci., Appl. Math. Appl. -ICCSAMA 2015 (2015) 279–289.

[18] A.P. Jain and P. Dandannavar, Application of machine learning techniques to sentiment analysis," in Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology, iCATccT 2016, Apr. 2017, pp. 628–632. doi: 10.1109/ICATCCT.2016.7912076.

[19] A. Naresh, P. Venkata Krishna, An efficient approach for sentiment analysis using machine learning algorithm, Evol. Intell. vol. 14 (2) (2021) 725–731, https://doi.org/10.1007/s12065-020-00429-1.

[20] A. Rahman, M.S. Hossen, Sentiment analysis on movie review data using machine learning approach (Sep), 2019 Int. Conf. Bangla Speech Lang. Process., ICBSLP 2019 (2019), https://doi.org/10.1109/ICBSLP47725.2019.201470.

[21] T. Susnjak, Applying BERT and ChatGPT for Sentiment Analysis of Lyme Disease in Scientific Literature," arXiv Prepr., no. arXiv:2302.06474, Feb. 2023, (Accessed: 6 June 2023). [Online]. Available: https://arxiv.org/abs/2302.06474v1.

[22] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao, Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-tuned BERT, arXiv Prepr., Feb. 2023, (Accessed: 6 June 2023). [Online]. Available: https://arxiv.org/abs/2302.10198v2.

[23] C. Leiter et al., ChatGPT: A Meta-Analysis after 2.5 Months, arXiv:2302.13795, Feb. 2023, (Accessed: 6 June 2023). [Online]. Available: http://arxiv.org/abs/2302.13795.

[24] J. Kocoń, et al., ChatGPT: Jack of all trades, master of none, Inf. Fusion (2023), 101861, https://doi.org/10.1016/J.INFFUS.2023.101861.

[25] J. Wang *et al.*, Is ChatGPT a Good NLG Evaluator? A Preliminary Study," arXiv: 2304.04339, Apr. 2023, (Accessed: 11 June 2023). [Online]. Available: https://arxiv.org/abs/2304.04339v1.

[26] Q. Xie, W. Han, Y. Lai, M. Peng, and J. Huang, The Wall Street Neophyte: A Zero-Shot Analysis of ChatGPT Over MultiModal Stock Movement Prediction Challenges," arXiv:2304.05351, Apr. 2023, (Accessed: 11 June 2023). [Online]. Available: https://arxiv.org/abs/2304.05351v2.

[27] Y. Zhu, P. Zhang, E.-U. Haq, P. Hui, and G. Tyson, Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks," arXiv:2304.10145, Apr. 2023, (Accessed: 11 June 2023). [Online]. Available: https://arxiv.org/abs/2304.10145v2.

[28] S. Haque, Z. Eberhart, A. Bansal, C. McMillan, Semantic similarity metrics for evaluating source code summarization,. Proc. 30th IEEE/ACM Int. Conf. Prog. Compreh., ICPC 2022, 2022, pp. 36–47, https://doi.org/10.1145/3524610.3527909.

[29] W. Zhang, Y. Deng, B. Liu, S.J. Pan, and L. Bing, Sentiment Analysis in the Era of Large Language Models: A Reality Check," arXiv:2305.15005, May 2023, (Accessed: 11 June 2023). [Online]. Available: https://arxiv.org/abs/2305.15005v1.

[30] A. Golubev, N. Rusnachenko, N. Loukachevitch, RuSentNE-2023: evaluating entity-oriented sentiment analysis on russian news texts, arXiv:2305. 17679 (2023), https://doi.org/10.28995/2075-7182-2022-20-XX-XX.

[31] M.U. Haque, I. Dharmadasa, Z.T. Sworna, R.N. Rajapakse, and H. Ahmad, I think this is the most disruptive technology': Exploring Sentiments of ChatGPT Early Adopters using Twitter Data," arXiv:2212.05856, Dec. 2022, (Accessed 25 October 2023). [Online]. Available: https://arxiv.org/abs/2212.05856v1.

[32] R.A. Laksono, K.R. Sungkono, R. Sarno, C.S. Wahyuni, Sentiment analysis of restaurant customer reviews on tripadvisor using naïve bayes, *Proc.* 2019 Int. Conf. Inf. Commun. Technol. Syst., ICTS 2019 (2019) 49–54, https://doi.org/10.1109/ICTS.2019.8850982.

[33] A. Soumeur, M. Mokdadi, A. Guessoum, A. Daoud, Sentiment Analysis of Users on Social Networks: Overcoming the challenge of the Loose Usages of the Algerian Dialect, Procedia Comput. Sci. vol. 142 (2018) 26–37, https://doi.org/10.1016/j.procs.2018.10.458.

[34] A.M. Alayba, V. Palade, M. England, and R. Iqbal, Arabic language sentiment analysis on health services, pp. 114–118, 2017, doi: 10.1109/asar.2017.8067771.

[35] A. Bhuvaneswari, J.T. Jones Thomas, P. Kesavan, Embedded Bi-directional GRU and LSTMLearning models to predict disasterson twitter data, Procedia Comput. Sci. vol. 165 (March) (2019) 511–516, https://doi.org/10.1016/j.procs.2020.01.020.

[36] M.A. Tocoglu, O. Ozturkmenoglu, A. Alpkocak, Emotion analysis from turkish tweets using deep neural networks, IEEE Access vol. 7 (2019) 183061–183069, https://doi.org/10.1109/ACCESS.2019.2960113.

[37] M. Kamyab, G. Liu, M. Adjeisah, Attention-Based CNN and Bi-LSTM model based on TF-IDF and GloVe word embedding for sentiment analysis, Appl. Sci. vol. 11 (23) (. 2021), https://doi.org/10.3390/app112311255.