information

Article

Analyzing Sentiments Regarding ChatGPT Using Novel BERT: A Machine Learning Approach

Sudheesh R [1,†], Muhammad Mujahid [2,†], Furqan Rustam [3,†], Rahman Shafique [4], Venkata Chunduri [5], Mónica Gracia Villar [6,7,8], Julién Brito Ballester [6,9,10], Isabel de la Torre Diez [11,*] and Imran Ashraf [4,*]

1 Kodiyattu Veedu, Kollam, Valakom 691532, India; sudheeshrofficial@gmail.com
2 Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan; mujahidws890@gmail.com
3 School of Computer Science, University College Dublin, D04 V1W8 Dublin, Ireland; furqan.rustam1@gmail.com
4 Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea; rahmanshafique47@gmail.com
5 Indiana State University, Terre Haute, IN 47809, USA; chunduriv1@gmail.com
6 Faculty of Social Science and Humanities, Universidad Europea del Atlántico, Isabel Torres 21, 39011 Santander, Spain; monica.gracia@uneatlantico.es (M.G.V.); julien.brito@uneatlantico.es (J.B.B.)
7 Department of Project Management, Universidad Internacional Iberoamericana Arecibo, Puerto Rico, PR 00613, USA
8 Department of Extension, Universidade Internacional do Cuanza, Cuito EN250, Bié, Angola
9 Universidad Internacional Iberoamericana, Campeche 24560, Mexico
10 Universitaria Internacional de Colombia, Bogotá 11001, Colombia
11 Department of Signal Theory, Communications and Telematics Engineering, Unviersity of Valladolid, Paseo de Belén, 15, 47011 Valladolid, Spain
* Correspondence: isator@tel.uva.es (I.d.l.T.D.); imranashraf@ynu.ac.kr (I.A.)
† These authors contributed equally to this work.

Abstract: Chatbots are AI-powered programs designed to replicate human conversation. They are capable of performing a wide range of tasks, including answering questions, offering directions, controlling smart home thermostats, and playing music, among other functions. ChatGPT is a popular AI-based chatbot that generates meaningful responses to queries, aiding people in learning. While some individuals support ChatGPT, others view it as a disruptive tool in the field of education. Discussions about this tool can be found across different social media platforms. Analyzing the sentiment of such social media data, which comprises people's opinions, is crucial for assessing public sentiment regarding the success and shortcomings of such tools. This study performs a sentiment analysis and topic modeling on ChatGPT-based tweets. ChatGPT-based tweets are the author's extracted tweets from Twitter using ChatGPT hashtags, where users share their reviews and opinions about ChatGPT, providing a reference to the thoughts expressed by users in their tweets. The Latent Dirichlet Allocation (LDA) approach is employed to identify the most frequently discussed topics in relation to ChatGPT tweets. For the sentiment analysis, a deep transformer-based Bidirectional Encoder Representations from Transformers (BERT) model with three dense layers of neural networks is proposed. Additionally, machine and deep learning models with fine-tuned parameters are utilized for a comparative analysis. Experimental results demonstrate the superior performance of the proposed BERT model, achieving an accuracy of 96.49%.

Keywords: ChatGPT; sentimental analysis; BERT; machine learning; LDA; app reviewers; deep learning

check for updates

Citation: R, S.; Mujahid, M.; Rustam, F.; Shafique, R.; Chunduri, V.; Villar, M.G.; Ballester, J.B.; Diez, I.d.l.T.; Ashraf, I. Analyzing Sentiments Regarding ChatGPT Using Novel BERT: A Machine Learning Approach. *Information* **2023**, *14*, 474. https://doi.org/10.3390/info14090474

Academic Editor: Peter Revesz

Received: 11 July 2023
Revised: 15 August 2023
Accepted: 19 August 2023
Published: 25 August 2023

1. Introduction

AI-based chatbots, powered by natural language processing (NLP), are computer programs designed to simulate human interactions by understanding speech and generating human-like responses [1]. They have gained popularity across various industries as a

tool to enhance digital experiences. The utilization of chatbots is experiencing continuous growth, with predictions indicating that the chatbot industry is expected to reach a market size of $3.62 billion by 2030, accompanied by an annual growth rate of 23.9% [2]. Additionally, the chatbot market is expected to reach approximately 1.25 billion U.S. dollars by 2025 [3]. The adoption of chatbots in sectors such as education, healthcare, banking, and retail is estimated to save around $11 billion annually by 2023 [4]. Especially in recent developments in the field of education, chatbots have the potential to significantly enhance the learning experience for students.

ChatGPT, an AI-based chatbot that is currently gaining attention, is being discussed widely across various platforms [5–7]. It has become a prominent topic of conversation due to its ability to provide personalized support and guidance to students, contributing to an improved academic performance. Developed by OpenAI, ChatGPT utilizes advanced language generation techniques based on the GPT language model technology [8]. Its impressive capabilities in generating coherent and contextually relevant responses have captivated individuals, communities, and social media platforms. The widespread discussions surrounding ChatGPT highlight its significant impact on natural language processing and artificial intelligence, and its potential to revolutionize our interactions with AI systems. People are fascinated by its usefulness in various domains including learning, entertainment, and problem-solving, which further contributes to its popularity and widespread adoption.

While there are many advantages to using ChatGPT, there are also some notable disadvantages and criticisms of the AI chatbot. Some raised concerns include the potential for academic dishonesty, as ChatGPT could be used as a tool for cheating in educational settings, similar to using search engines like Google [9]. There is also a concern that ChatGPT may perpetuate biases when used in research, as the language model is trained on large amounts of data that may contain biased information [9]. Another topic of discussion revolves around the potential impact of ChatGPT on students' critical thinking and creativity. Some argue that an over-reliance on ChatGPT may lead to a decline in these important skills among students [10]. Additionally, the impact of ChatGPT on the online education business has been evident, as seen in the case of Chegg Inc., where the rise of ChatGPT contributed to a significant decline of 47% in the company's shares during early trading [11]. To gain insights into people's perceptions of ChatGPT, opinion mining was conducted using social media data. This analysis aimed to understand the general sentiment and opinions surrounding the use of ChatGPT in various contexts: people, in this sense, tweet on Twitter concerning their thoughts about ChatGPT, which could provide valuable information.

Opinion mining involves evaluating individuals' perspectives, attitudes, evaluations, and emotions towards various objects including products, services, companies, individuals, events, topics, occurrences, and applications, along with their attributes. When making decisions, we often seek the opinions of others, whether as individuals or organizations. Sentiment analysis tools have found application in diverse social and corporate contexts [12]. Social media platforms, microblogging sites, and app stores serve as rich sources of openly expressed opinions and discussions, making them valuable for a sentiment analysis [13]. The sentiment analysis employs NLP, a text analysis, and computational methods such as machine learning and data mining to automate the categorization of sentiments based on feedback and reviews [14]. The sentiment analysis process involves identifying sentiment from reviews, selecting relevant features, and performing sentiment classification to determine polarity.

1.1. Research Questions

To meet the objective of this study by analyzing people's attitudes toward ChatGPT, this study formulates the following questions (RQs):

i. **RQ1:** What are people's sentiments about ChatGPT technology?

ii. **RQ2:** Which classification model is most effective, such as the proposed transformer-based models, machine learning-based models, and deep learning-based models, for analyzing sentiments about ChatGPT tweets?
iii. **RQ3:** What are the impacts of ChatGPT on student learning?
iv. **RQ4:** What role does topic modeling play in the sentiment analysis of social media tweets?

1.2. Contributions

The sentiment analysis of tweets regarding ChatGPT aims at providing users' perceptions of ChatGPT and analyzing the ratio of positive and negative comments from users. In addition, a topic analysis can provide insights on frequently discussed topics concerning ChatGPT and provide feedback to further improve its functionality. In particular, the following contributions are made:

- This study aims to analyze people's perceptions of the trending topic of ChatGPT worldwide. The research contributes by collecting relevant data and examining the sentiments expressed by individuals toward this significant development.
- Tweets related to ChatGPT are collected by utilizing the Tweepy application programming interface (API) and employing various keywords. The collected tweets undergo preprocessing and annotation using Textblob and the valence aware-dictionary (VADER). The bag of words (BoW) feature engineering technique is employed to extract essential features.
- A deep transformer-based BERT model is proposed for the sentiment analysis. It consists of three dense layers of neural networks for enhanced performance. Additionally, machine learning and deep learning models with fine-tuned parameters are utilized for comparison purposes. Notably, this study is the first to investigate ChatGPT raw tweets using Transformers.
- The study utilizes the latent Dirichlet allocation (LDA) approach to extract highly discussed topics from the dataset of ChatGPT tweets. This analysis provides valuable insights into the frequently discussed themes and subjects.

The remaining sections of the paper are structured as follows: Section 2 provides a comprehensive review of relevant research works on sentiment analyses, offering a valuable background for the proposed approach. Section 3 presents a detailed description of the proposed approach. Section 4 presents and discusses the experimental results obtained from the analysis. Finally, Section 5 concludes the study, summarizing the key findings and suggesting potential directions for future research.

2. Related Work

The analysis of reviews has gained significant attention in recent years, mainly due to the widespread use of social media platforms. These platforms serve as a hub for discussions on various topics, providing researchers with valuable insights and information. For instance, in a study conducted by Lee et al. [15], social media data were utilized to investigate the Taliban's control over Afghanistan. By analyzing the discussions and conversations on social media, the study aimed to gain a deeper understanding of the situation. Similarly, the study by Lee et al. [16] focused on extracting tweets related to racism to shed light on the issue of racism in the workplace. By analyzing these tweets, the researchers aimed to uncover patterns and gain insights into the prevalence and nature of racism in professional environments. They utilized Twitter data and annotated it with the TextBlob approach. The authors attained 72% accuracy for the racism classification. In a different context, Mujahid et al. [17] conducted a study on public opinion about online education during the COVID-19 pandemic. By analyzing social media data, the researchers aimed to understand the sentiment and perceptions surrounding online education during this challenging time. These studies highlight the significance of a social media data analysis in extracting meaningful information and gaining insights into various subjects. By harnessing the vast amount of discussions and conversations on social media platforms,

researchers can delve into important topics and uncover valuable findings. The researchers employed 17,155 tweets for the analysis and attained 95% accuracy using the SMOTE technique with bag of word features by the SVM model.

ChatGPT is a hot topic nowadays and exploring people's perceptions about it using Twitter data can provide valuable insights. Many studies have previously done such kinds of analyses on different topics. In the study conducted by Tran et al. [18], the focus was on examining consumer sentiments towards chatbots in various retail sectors and investigating the impact of chatbots on their sentiments and expectations regarding interactions with human agents. Through the application of the automated sentiment analysis, it was observed that the general sentiment towards chatbots is more positive compared to that towards human agents in online settings. They collected a limited dataset of 8190 tweets and used ANCOVA for the test. They only classify the tweets into their exact sentiments and do not properly use performance metrics like accuracy. Additionally, sentiments varied across different sectors, such as fashion and telecommunications, with the implementation of chatbots resulting in more negative sentiments towards human agents in both sectors. The study [19] aimed to develop an effective system for analyzing and extracting sentiments and mental health during the COVID-19 pandemic. By utilizing a vast amount of data and leveraging hashtags, we employed the BERT machine learning algorithm to classify customer perspectives into positive and negative sentiments with high accuracy. Ensuring user privacy, our main objective was to facilitate self-understanding and the regulation of mental states through end-to-end encrypted user-bot interactions. The researchers were able to achieve 95.6% accuracy and 95% recall for automated sentiment classification related to chatbots.

Some studies, such as [20], focus on a sentiment analysis of disaster-related tweets at different time intervals for specific locations. By using the LSTM network with word embedding, keywords are derived from the tweet history and context. The proposed algorithm, RASA, classifies tweets and identifies sentiment scores for each location. RASA outperforms other algorithms, aiding the government in post-disaster management by providing valuable insights and preventive measures. Another study [21] tries to predict cryptocurrency prices using Twitter data. They focus on a sentiment analysis and emotion detection using tweets related to cryptocurrency. An ensemble model, LSTM-GRU, combines LSTM and GRU to enhance the analysis' accuracy. Multiple features and models, including machine learning and deep learning, are examined. Results reveal a predominance of positive sentiment, with fear and surprise also as prominent emotions. The dataset consists of five emotions extracted from Twitter. The proposed ensemble model achieves 83% accuracy using a balanced dataset for emotion prediction. This research provides valuable insights into the public perception of cryptocurrency and its market implications.

Additionally, it is also observed that most of the time, a service provider asks for feedback regarding the quality or satisfaction level of the services or products via a customer feedback form provided in an online mode, most probably by using a social media platform [22]. Such assessments are critical in determining the quality of services and products. However, it is necessary to examine the views of user concepts and impressions. Negative sentiment ratings, in particular, include more relevant recommendations for enhancing the quality of the product/service. Given the significance of the text analysis, there is a huge amount of work on the sentiment analysis. For example, studies [23–25] classify app reviews by using machine learning and deep learning models. Another piece of research [26] looked at the Shopify app reviews and classified them as pleased or dissatisfied. For sentiment classification, many feature extraction approaches are used in conjunction with supervised machine learning algorithms. For the experiments, 12,760 samples of app reviews were utilized with machine learning. Different hybrid approaches to combining the features were used to enhance the performance. But LR performed with 83% accuracy and an 86% F score. The performance of machine learning models in the sentiment analysis can be influenced by the techniques used for feature engineering. Research studies [27,28] indicate that altering the feature engineering process can result in

changes to the models' performance. The research [29] provides a method for categorizing and evaluating employee reviews. For employee review classification, it employs an ETC with BoW features. The study classified employee reviews using both numerical and text elements and achieved 100% and 79% accuracy, respectively. Ref. [30] used NB in conjunction with the RF and SVM to categorize mobile app reviews from the Google Play store. The researcher collected over 90,000 reviews posted in the English language for 10 applications available on the Google Play Store. A total of 7500 reviews were annotated from a dataset of 90,000 tweets. The final experiments implemented the use of 7500 reviews. The results indicated that a baseline 10-fold validation yielded an accuracy of 90.8%. Additionally, the precision was found to be 88%, the recall was 91%, and the f score was 89%. Ref. [31] also used an RF algorithm to identify the variables that distinguish reviews from those from other nations. The research [32] looked at retail applications in Bangladesh. The authors gathered data from Google Play and utilized VADER and AFFIN to annotate sentiments. For sentiment categorization, many machine learning models are employed, and RF outperforms with substantial accuracy. Bello et al. [33] proposed a BERT model for a sentiment analysis on Twitter data. The authors used the BERT model with different variants including the recurrent neural network (RNN) and Bi-long short-term memory (BILSTM) for classification. Catelli et al. [34] and Patel et al. [35] also employed the BERT model for a sentiment analysis on app reviews with lexicon-based approaches.

The study [36] presented a hybrid approach for the sentiment analysis of ChatGPT tweets. Raw tweets were transformed into structured and normalized forms to improve the accuracy of the model and a lower computing complexity. For the objective of classifying tweets from ChatGPT, the authors developed hybrid models. Although state-of-the-art models are unable to provide correct predictions, hybrid models incorporate multiple models to eliminate bias, improve overall outcomes, and make precise predictions. Bonifazi et al. [37] proposed a framework for determining the spatial and spatio-temporal extent of a user's sentiment regarding a topic on a social network. First, the authors introduced the idea of their research, observing that it summarizes a number of previously discussed ideas about social websites. In reality, each of these ideas represents a unique fact about the concept. Then, they established a framework capable of expressing and controlling a multidimensional view-of scope, which is the sentiment of an individual regarding a topic. After that, they recommended a number of parameters and a method for assessing the spatial and spatio-temporal scope of a user's opinion on a topic on a social platform. They conducted several experiments on actual data collected through Reddit to test the proposed framework. Similarly, Bonifazi et al. [38] presented another Reddit-based study. They proposed a model for evaluating and visualizing the eWoM Power of Reddit blog posts.

In a similar way, ref. [39] examined app reviews, where the authors initially extracted negative reviews, constructed a time series of these reviews, and subsequently trained a model to identify key patterns. Additionally, the study focused on an automatic review classification to address the challenge of handling a large volume of daily submitted reviews. To tackle this, the study presented a multi-label active-learning technique, which yielded superior results compared to state-of-the-art methods. Given the impracticality of manually analyzing a vast number of reviews, many researchers have turned to topic modeling, a technique that aids in identifying the main themes within a given text. For instance, in the study [40], the authors investigated the relationship between Arabic app elements and assessed the accuracy of reflecting the type and genre of Arabic mobile apps available on the Google Play store. By employing the LDA approach, valuable insights were provided, offering potential improvements for the future of Arabic apps. Furthermore, in [41], the authors developed an NB and XGB technique to determine user activity within an app.

The literature review provides an analysis of the advantages, disadvantages, and limitations associated with different approaches. Nevertheless, it is worth noting that a significant number of researchers have directed their attention toward the utilization of Twitter datasets for the purpose of analyzing tweets and app evaluations. The researchers employed natural language processing (NLP) techniques and machine learning primarily

for the purpose of a sentiment analysis. Commonly utilized Machine learning models, including random forests, support vector machines, and extra tree classifiers, are limited in their ability to learn intricate patterns and are typically not utilized for large datasets. When the aforementioned models are employed on extensive datasets, their performance is inadequate and demands an excessive amount of time for training, especially in the case of handcrafted features. Furthermore, the existing literature employs a limited collection of datasets, which are only comprised of tweets that are not linked to ChatGPT tweets. Previous research has not extensively examined the topic of ChatGPT or OpenAI-related tweets and achieved a low accuracy. Table 1 shows the summary of the literature review.

Table 1. Summary of related work.

Authors	Techniques	Advantages	Disadvantages	Limitations
[16]	TextBlob, CNN, RNN, GRU, DT, RF, SVM	The authors make an ensemble model by combining the GRU, CNN, and RNN for the extraction of features from the tweets and detection. They also performed seven experiments to test the proposed ensemble approach.	The authors develop ensemble models, which need a significant amount of time to both train and identify the sentiments.	The authors used a limited dataset and did not develop transformer-based models that are the most up-to-date and that provide high accuracy.
[17]	TextBlob, CNN, LSTM, SVM, GBM, KNN, DT, LSTM-CNN	This study employed machine learning as well as deep learning for the analysis of tweets. They utilized various annotation and feature engineering techniques. Machine learning outperformed deep learning with an accuracy of 95%.	The study did not clearly describe the preprocessing stages and their implementations.	The dataset included in this study was restricted to tweets that were not associated with ChatGPT tweets.
[18]	BERT	The authors conducted this research to analyze the depression tweets during the period of COVID-19 and achieved remarkable results with BERT.	To speed up computation, the research did not remove stopwords, punctuation, numerical values, etc., from the text. Additionally, the accuracy was inadequate.	The research only proposed one model, which was BERT, and did not compare with other studies.
[19]	Naïve Bayes	The data in the study was labeled using the Vader technique, and the Nave Bayes model was implemented to examine the influence of chatbots on customer opinions and demands within the retail industry.	The study detected positive, neutral, and negative sentiments and used the Ancova test only for the experiments.	The study did not use the most important metrics like accuracy, deep models, or transformers. The study is limited to the Nave Bayes model.
[21]	LSTM + GRU, CNN, SVM, DT, TFIDF	Their primary area of research revolves around sentiment evaluation and detecting emotions using tweets that are associated with cryptocurrencies. The utilization of an ensemble model, namely the LSTM-GRU model, involves the integration of both LSTM and GRU architectures in order to improve the accuracy of the analysis.	The author used ensemble models, which necessitate substantial time for both training and sentiment identification.	The study is regarding the cryptography analysis. Also, transformers are ignored in this study.
[26]	RF, LR, and AC	The study used various feature engineering strategies, including bag-of-words; term frequency, inverse document-frequency, and Chi-2 are employed individually and collectively in order to attain meaningful information from the tweets.	The study employed various feature engineering strategies but did not use cross-dataset experiments with machine learning classifiers. The LR achieved a 83% lowest accuracy.	The study does not use Chatbots or ChatGPT-related tweets for the experiments. In addition, their focus is on utilizing machine learning models for Shopify reviews.
[30]	SVM, RF, and NB	The dataset was obtained by the authors from the most popular ten applications. The findings of the study revealed that a baseline 10-fold validation approach resulted in an accuracy rate of 90.8%.	The paper is about app reviews, not ChatGPT tweets.	The accuracy achieved is very low, and the study did not use any deep transformers to improve its efficiency.

As a result, this paper proposes a transformer-based BERT model that leverages self-attention mechanisms, which have demonstrated remarkable efficacy in the context of machine learning and deep learning. The proposed model addresses the problems mentioned in the literature review. They have the ability to comprehend the correlation between consecutive items that are widely separated. The transformers achieved an exceptional performance. Additionally, the performance of the proposed method was

evaluated using cross-validation findings and statistical tests. The ChatGPT tweets study utilizes BERTopic and LDA-based topic modeling techniques to ascertain the most pertinent topics or keywords within the datasets.

3. Methodology

The proposed methodology's workflow is depicted in Figure 1, illustrating the steps involved. Firstly, unstructured tweets related to ChatGPT are collected from Twitter using the Twitter Tweepy API. These tweets undergo several preprocessing steps to ensure cleanliness and remove noise. Lexicon-based techniques are then utilized to assign labels of positive, negative, or neutral to the tweets. Feature extraction is performed using the Bag of Words (BoW) technique on the labeled dataset. The data is subsequently split into an 80/20 ratio for training and testing purposes. Following model training, evaluation metrics such as accuracy, precision, recall, and the F1 score are employed to analyze the model's performance. Each component of the proposed methodology for sentiment classification is discussed in greater detail in the subsequent sections.

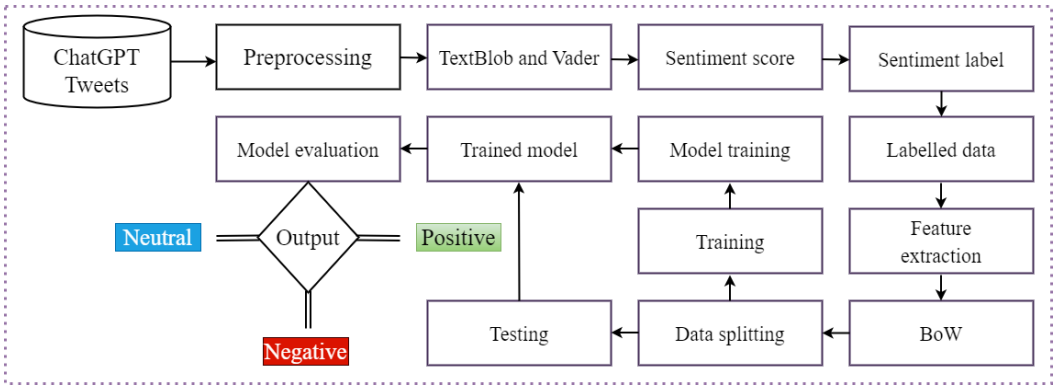

Figure 1. The workflow diagram of the proposed approach for sentiment classification.

3.1. Dataset Description and Preprocessing

In this study, the ChatGPT tweets dataset is utilized, which is scraped from Twitter using the Tweepy API Python library. A total of 21,515 raw tweets are collected for this purpose. The dataset contains the date, user name, user friends, user location, and text features. The dataset is unstructured and requires several preprocessing steps to make it appropriate for machine learning models.

Text preprocessing is very important in NLP tasks for a sentiment analysis. The dataset used in this paper is unstructured, unorganized, and contains unnecessary and redundant information. The machine learning or deep learning models do not perform well on these types of datasets, which increases the computational cost [42]. Different preprocessing techniques are utilized to remove unnecessary, meaningless information from the tweets. Preprocessing is a crucial step in data analysis that involves transforming unstructured data into a meaningful and comprehensible format [43]. The purpose of preprocessing is to enhance the quality of the dataset while preserving its original content, enabling the model to identify significant patterns that can be utilized to extract valuable and efficient information from the preprocessed data. There are many steps in preprocessing to convert unstructured text into structured data. These techniques are used to remove the least important information from the data and make it easier for the machine to train in less time.

The dataset consists of 20,801 tweets, 8095 of which are positive, 2727 of which are negative, and 9979 of which are neutral. Following the split, 6476 positive tweets were used for training and 1619 for testing. There were 1281 negative tweets utilized for training and 546 for testing. For neutral tweets, 7983 were training and 1996 were testing. The hashtags #chatgpt, #ChatGPT, #OpenAI, #ChatGPT-3, #Chatbots, #Powerful OpenAI, etc., were used to collect all of the tweets in English. Table 2 shows the dataset statistics.

Table 2. Dataset statistics after splitting.

Tweets	Training	Testing	Total
Positive	6476	1619	8095
Negative	1281	546	2727
Neutral	7983	1996	9979
Total	16,640	4161	20,801

The most important step in natural language processing (NLP) is the pre-processing stage. It enables us to remove any unnecessary information from our data so that we can proceed to the following processing stage. The Natural Language Toolkit (NLTK), which provides modules, is an open-source Python toolkit that can be used to perform operations such as tokenization, stemming, classification, etc. The first step in preprocessing is to convert all textual data to lowercase. Conversion is an essential step in sentiment classification, as the machine considers "ChatGPT" and "chatgpt" as individual words. The dataset contains text in upper, lower, and sentence case, which the model takes separately, which affects the classification performance as well and makes the data more complex if we do not convert it all into lowercase. The second step is to remove numbers from the text because they do not provide meaningful information and are useless in the decision-making process. The removal of numerical data enhances the quality of the data [44]. The third step is to remove punctuation such as [?,@,#,/,&,%] to increase the quality of the dataset and the performance of the models. The fourth step is to remove HTML and URL tags that also provide no important information. The URLs in the text data are meaningless because they expand the dataset and require extra computation. It has no impact on the machine learning performance. The fifth step is to remove stopwords like 'an', 'the', 'are', 'was', 'has', 'they', etc., from the tweets during preprocessing. The model's accuracy improves, and the training process is faster, with only relevant information [44]. Additionally, the removal of stopwords allows for a more thorough analysis, which is advantageous for a limited dataset [45]. The last step is to perform stemming and lemmatization. The effectiveness of machine learning is slightly influenced by the stemming and lemmatization steps. After performing all important preprocessing steps, the sample tweets are presented in Table 3.

Table 3. Sample Tweets before preprocessing and after preprocessing.

Unstructured Tweets	Structured Tweets (Preprocessed)
I asked #chatgpt to write a story instalment with Tim giving the octopus a name. Originality wasn't its strongpoint €| https://t.co/rbB5prcJ2r (accessed on 2 April 2023).	asked chatgpt write story instalment tim giving octopus name originality strongpoint
ChatGPT is taking the web by storm; If you're unable to try it on their site, feel free to test it out through us! €| https://t.co/jfmOQmjSHo (accessed on 2 April 2023).	chatgpt taking web storm unable try site feel free test
People weaponizing a powerful AI tool like ChatGPT days into launch has to be the most predictable internet	people weaponizing powerful tool like chatgpt days launch predictable internet

3.2. Lexicon Based Techniques

TextBlob [46] and VADER [47] are the two most important lexicon-based techniques used in this study to label the dataset. TextBlob provides the subjectivity and polarity scores, where 1 represents the positive response and -1 represents the negative response in polarity. The subjectivity score is represented by [0, 1]. The VADER technique calculates the sentiment score by adding the intensity of each word in the preprocessed text.

3.3. Feature Engineering

The labeled dataset is divided into training and testing subsets. The training data has been used to fit the model, while the test data is used by the model for predictions on unseen data, which are then compared to determine the model's efficacy.

Important features from the cleaned tweets are extracted using the BoW approach. The BoW approach extracts valuable features from the data to enhance the performance of machine learning models. Features are very crucial and have a great impact on sentiment classification. This approach reduces processing time and effort. The BoW approach creates a bag of words of text data and converts it into a numeric format. The models learn and understand complex patterns and sequences from the numeric format [48].

3.4. Machine and Deep Learning Models

This subsection provides details about the machine and deep learning models. The applications of machine and deep learning span across various domains, such as disease diagnosis [49], education [50], computer/machine vision [51,52], text classification [53], and many more. In this study, we utilize these techniques for text classification. The objective of text classification is to automatically classify texts into predetermined categories. Deep learning and machine learning are both forms of artificial intelligence [54]. Classification of text using machine learning entails the transformation of input data into a numeric form. Then, manually extracting features from the data using a bag of words, term frequency, inverse document frequency, word2vec, etc., to extract crucial features. Frequently employed models of machine learning, such as random forests, support vector machines, extra tree classifiers, etc., cannot learn complex patterns and are not employed for large datasets. When we apply these models to large datasets, they perform poorly and require excessive training time, particularly for handcrafted features. If the researchers applied machine learning to complex problems, they would require manual feature engineering to retain only the essential information, which is time-consuming and requires expertise in the same fields to improve classification results.

Deep learning [55], on the other hand, has a method for automatically extracting features. Large and complex patterns are automatically learned from the data using DL models like CNN, LSTM, GRU, etc., minimizing the need for manual feature extraction. When there is a lack of data, the model could get overfitted and perform poorly. These models address the issue of vanishing gradients. In terms of computing, gated recurrent units (GRU) outperform LSTM, reduce the chances of overfitting, and are better suited for small datasets. Additionally, GRU has a straightforward structure with fewer parameters. The authors only used models that are quick and effective in terms of computing.

We developed transform-based models that use self-attention mechanisms since they are the most effective after machine and deep learning. They have the capacity to comprehend the relationship between consecutive elements set far apart from one another. They achieve an outclass performance. They give each component of the sequence the same amount of attention. The large data can be processed and trained by transformers in a shorter period of time. They are capable of processing almost any form of sequenced information. The hyperparameters and their fine-tuned values are represented in Table 4. These parameters are obtained using the GridSearchCV method which performs an exhaustive search for the given parameters to evaluate a model's performance and provides the best set of parameters for obtaining optimal results.

Table 4. Hyperparameters and their tuned values for experiments.

Model	Parameters Tuning
RF	n_estimators = 100, random_state = 50, max_depth = 150
GBM	n_estimators = 100, random_state = 100, max_depth = 300
LR	random_state = 150, solver = 'newton-cg', multi_class = 'multinomial', C = 2.0

Table 4. *Cont.*

Model	Parameters Tuning
SVM	kernel = 'linear', C = 1.0, random_state = 200
KNN	n_neighbors = 3
DT	random_state = 100, max_depth = 150
ETC	n_estimators = 100, random_state = 150, max_depth = 300
SGD	loss = "hinge", penalty = "l1", max_iter = 6
CNN	616,003 trainable parameters
RNN	633,539 trainable parameters
LSTM	655,235 trainable parameters
BILSTM	726,787 trainable parameters
GRU	692,547 trainable parameters

- Logistic Regression: LR [56] is a simple machine learning model used in this study for sentiment classification. LR provides accurate results with preprocessed and highly relatable features. It is simple to implement and utilizes low computational resources. This model may not perform well on large datasets, cause overfitting, and does not learn complex patterns due to its simplicity.
- Random Forest: The RF is an ensemble supervised machine learning model used for classification, regression, and other NLP tasks [57]. The RF ensembles multiple decision trees to form a forest. A large amount of textual data and the ensemble of trees make the model more complex which takes a higher amount of time to train. The RF is powerful and has attained high accuracy for the sentiment analysis.
- Decision Tree: A DT is a supervised non-parametric learning model for classification and regression. The DT predicts a target variable using learned features to classify objects. A decision tree requires less data cleaning than other machine learning methods. In other words, decision trees do not require normalization during the early stages of machine learning tasks. They can handle both categorical and numerical information [58].
- K Nearest Neighbour: The KNN model requires no previous knowledge and does not learn from training data. It is also called the lazy learner. It does not perform well when data is not well normalized and structured. The performance can be manipulated with the distance metrics and K value [59].
- Support Vector Machine: The SVM is mostly used for classification tasks. It performs well where the number of dimensional spaces is greater than the number of samples [17]. The SVM does not perform well on large datasets because the training time increases. It is more robust and handles imbalanced datasets efficiently. The SVM can be used with 'poly', 'linear', and 'rbf' kernels.
- Extra Tree Classifier: The ETC is used for classification and regression [60]. Extra trees do not use the bootstrapping approach and train faster. The ETC requires fewer parameters for tuning compared to RF. Also, with extra trees, the chances of overfitting are less.
- Gradient Boosting Machine (GBM) and Stochastic Gradient Descent (SGD): The GBM [61] and SGD are supervised learning models for classification. To enhance the performance, the GBM combines multiple decision trees, and the SGD optimizes the gradient descent. The GBM is more complex and handles imbalanced data better than the SGD.
- Convolutional Neural Networks (CNN): The CNN [62] is a deep neural network model that is used for image classification, sentiment classification, object detection, and many other tasks. For sentiment classification, it first converts textual data into a numeric format, then make a matrix of word embedding layers. These embedding

layers are then passed into convolutional, max-pooling, and dense layers, and the final output is passed through a dense softmax layer for classification.

- Recurrent Neural Network (RNN): The RNN [63] is a widely used model for text classification, speech recognition, and NLP tasks. The RNN can handle sequential data with complex long-term dependencies. This model is expensive to train and has the vanishing gradient issue for text classification.
- Long Shor-Term Memory: The LSTM [64] model was released to handle long-term dependencies, the gradient vanishing issue, and the complex training time. When compared to RNN, this model is much faster and uses less memory. It has three gates, including input, output, and forget, which are used to manage the data flow.
- Bidirectional LSTM: The BiLSTM is a deep learning model which is used for several tasks, including text classification as well [65]. The model provides better results for understanding the text in past and future contexts than the LSTM. It can learn information from both directions.
- Gated Recurrent Unit (GRU): The GRU solves the problem of vanishing gradient, faced by RNN [66]. It is fast and performs well on small datasets. The model has two gates: an update gate and a reset gate.

3.5. Transformer Based Architecture

BERT is a transformer-based model presented by Devlin et al. [67] in 2018. The BERT model uses an attention mechanism that takes actual input from the text. The BERT has two parts: an encoder and a decoder. The encoder gets the input as text and produces output such as predictions. The BERT model is particularly well suited for NLP tasks, including a sentiment analysis and questioning-and-answering, because it is trained on a large amount of textual data. The traditional models only use word context-of-word in just one direction, normally from left to right. The BERT model considers the context of words in NLP in both directions. In contrast to previous deep learning models, this model has a clear understanding of word meanings. The BERT model is trained on a large amount of data to obtain accurate results and to learn complex patterns and structures [68].

The BERT with fine-tuned hyperparameters works well for a variety of NLP tasks. Santiago Gonzalez and Eduardo C. Garrido-Merchan [69] published a study that compared the BERT architecture to traditional machine learning models for sentiment classification. The traditional models were trained using features extracted from TF-IDF. The performances demonstrate that the BERT transformer-based model outperforms the traditional models. To solve NLP-related problems, the BERT model has also been used for low-resource languages. BERT was used to pre-train text data and fine-tuned low-resource languages by Jan Christian Blaise Cruz and Charibeth Cheng [70]. Because this model takes input words with multiple word sequences at once, the results for that language were improved.

Figure 2 shows the proposed architecture of BERT for sentiment classification. The BERT uses a large, pre-trained vocabulary to generate input ids that are numeric values of the input text. First of all, a sequence of tokens is created from whole input text tokens, and unique ids are assigned to the tokens. Basically, input ids are numerical representations of input text. In BERT, the input mask works like an attention mechanism, which clearly differentiates between input text tokens and padding. The input mask identifies which tokens in the input sequence are evaluated by the model and which ones are not evaluated. Segment ids indicate extra tokens to differentiate different sentences. After that, it is concatenated with the BERT Keras layer. This study uses three dense layers in BERT with 128, 64, and 32 units and two 20% dropout layers. The final dense layer is used for classification with the softmax activation function.

XLNet was released by Ashish Vaswani in 2019, and its architecture is similar to BERT. The BERT is an auto-encoder, and the XLNet is an autoregressor model [71]. The BERT model cannot correctly model the dependencies between tokens in a sentence. XLNet overcomes this problem by adopting permutation-based training objectives as compared to

mask-based objectives. The permutation-based objective permits XLNet to represent the dependencies with all tokens in a paragraph.

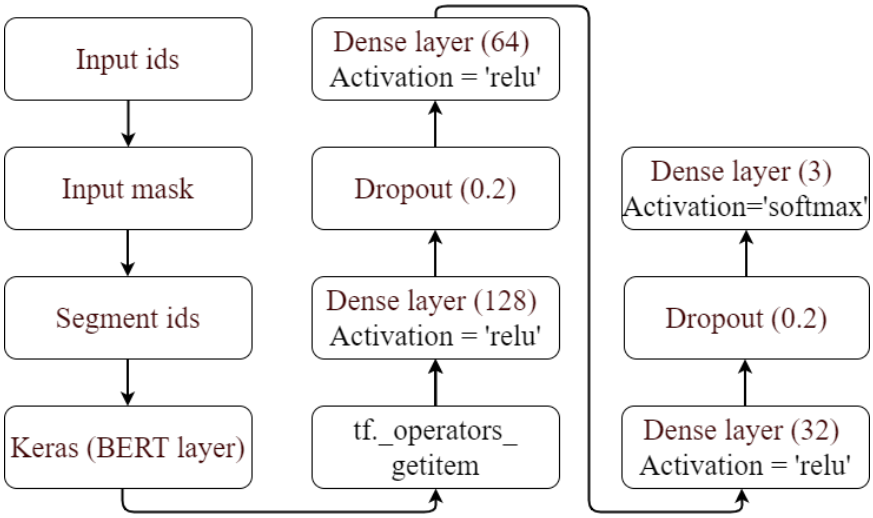

Figure 2. The architecture for the proposed sentiment classification.

Robustly optimized BERT pretraining (RoBERTa) [72] is a transformer-based model used for various NLP tasks. It was developed in 2019. RoBERTa is a modification of the BERT model to overcome the limitations of the BERT model. RoBERTa is trained on 160 billion words, whereas BERT is trained on only 3.3 billion words. RoBERTa is trained on large data sets, is fast to train, and may use large batch sizes. RoBERTa uses a dynamic masking approach, and BERT uses a static approach.

3.6. Performance Metrics

The performance of the machine, deep, and transformer-based models are also measured using evaluation metrics including accuracy, precision, recall, and the F1 score [73]. Accuracy is calculated using

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{1}$$

where TP stands for true positive, TN for true negative, FP for false negative, and FN for false negative.

Precision is another performance metric used to measure performance. Precision is defined as the ratio of actual positives to the total number of positive predictions.

$$Precision = \frac{TP}{(TP + FP)} \tag{2}$$

The recall is also used to measure the performance of models. The recall is calculated by dividing the true positives by the sum of true positives and false negatives.

$$Recall = \frac{TP}{(TP + FN)} \tag{3}$$

The F1 score is a better metric than other metrics in a situation where classes are imbalanced because it considers both precision and recall and provides a better understanding of the model's performance.

$$F1 - score = 2 * \frac{(Recall * precision)}{(Recall + precision)} \tag{4}$$

4. Results and Discussion

This section presents the details regarding experiments on the ChatGPT Twitter dataset using machine learning, deep learning, and transformer-based models. The Colab Notebook in Python with Tensorflow, Keras, and Sklearn libraries is used to evaluate the research experiments. Different measures including accuracy, precision, recall, and the F1 score are used to assess the performance of various models. For deep and transformer-based models, a graphics processing unit (GPU) and 16 GB of RAM are used to speed up the training process. Experimental results are presented in the subsequent sections.

4.1. Results of Machine Learning Models

Table 5 shows the results of eight machine learning models utilizing Textblob and VADER lexicon-based techniques on ChatGPT Twitter data. With an accuracy of 94.23%, SVM outperforms while SGD achieves an accuracy of 92.74%. A 91% accuracy is attained by ETC, GBM, and LR while the lazy learner KNN obtains only a 58.03% accuracy. The SVM model has 88% accuracy, 89% recall, and an 83% F1 score for the negative class, whereas the GBM model has 91% precision, 63% recall, and a 74% F1 score. Utilizing BoW features, the neutral tweets get the highest recall scores.

Table 5. Results of machine learning models using VADER and TextBlob techniques.

			Vader				**TextBlob**		
Model	**Accuracy**	**Class**	**Precision**	**Recall**	**F1 Score**	**Accuracy**	**Precision**	**Recall**	**F1 Score**
SGD	89.13	Positive	93	92	93	92.76	94	93	93
		Negative	84	69	76		89	75	81
		Neutral	87	94	90		93	95	97
RF	82.40	Positive	92	83	88	86.99	94	85	89
		Negative	92	43	58		94	47	63
		Neutral	73	98	84		82	99	90
DT	82.26	Positive	93	82	87	88.29	94	85	90
		Negative	82	47	60		89	56	69
		Neutral	94	97	84		84	99	91
ETC	87.11	Positive	93	89	91	91.80	94	91	93
		Negative	92	56	69		90	66	76
		Neutral	81	98	89		90	99	94
KNN	54.38	Positive	95	47	22	58.03	95	20	34
		Negative	83	20	33		80	18	30
		Neutral	47	99	64		54	99	70
SVM	90.72	Positive	95	92	94	94.23	96	94	95
		Negative	85	73	79		88	89	83
		Neutral	89	96	92		94	99	96
GBM	89.56	Positive	93	92	92	92.28	94	94	94
		Negative	92	65	76		91	63	74
		Neutral	85	97	91		91	99	95
LR	88.44	Positive	93	91	92	91.56	95	91	93
		Negative	89	63	74		92	66	77
		Neutral	84	96	90		89	99	96

Table 5 also shows the results of various models using the VADER technique. Using a VADER lexicon-based technique, SVM performs best with an accuracy of 90.72%. The models SGD and GBM both achieved an 89% accuracy score. The model that performs worse, in this case, is KNN, with a 54.38% accuracy. This model also performs poorly on the TextBlob technique. The only model in machine learning that performs with the highest accuracy is SVM with the linear kernel. The accuracy score of various machine learning models using TextBlob and Vader are compared in Figure 3.

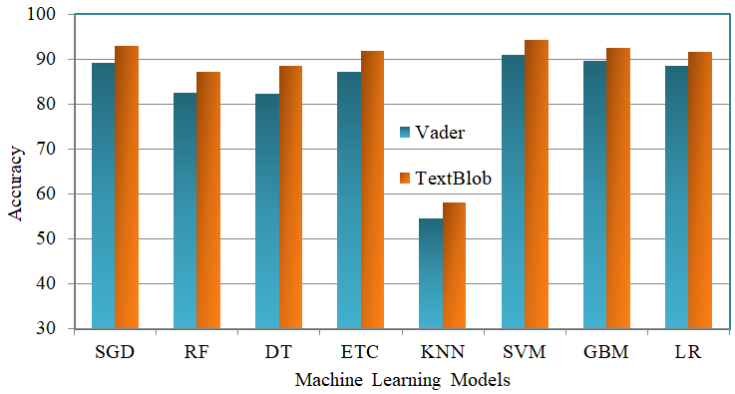

Figure 3. Performance of models using the TextBlob and VADER techniques. The X-axis presents the machine learning models that we utilized in this study, and the Y-axis presents the accuracy score.

4.2. Performance of Deep Learning Models

Deep learning models are also used to perform a sentiment classification and analysis. Results using the TextBlob technique are shown in Table 6. The experimental results on the ChatGPT preprocessed Twitter dataset show that the BiLSTM deep model achieves a 93.12% accuracy score, which is the highest as compared to CNN, RNN, LSTM, and GRU. The LSTM model also performs well, with an accuracy score of 92.95%. The other two deep models, GRU and RNN, reached an accuracy higher than 90%. The performance of the CNN model is not good. The CNN model achieved a 20% lower accuracy than other models.

Table 6. Results of deep learning models using the TextBlob technique.

Model	Accuracy	Class	Precision	Recall	F1 Score
CNN	70.88	Positive	73	66	69
		Negative	56	48	52
		Neutral	71	81	77
RNN	90.35	Positive	91	92	92
		Negative	80	71	75
		Neutral	92	94	93
LSTM	92.95	Positive	93	94	93
		Negative	83	82	82
		Neutral	96	96	96
BiLSTM	93.12	Positive	91	96	93
		Negative	86	81	83
		Neutral	97	94	12
GRU	92.33	Positive	92	94	93
		Negative	82	81	82
		Neutral	95	94	95

Table 7 shows the results of deep learning using the VADER technique. The performance of five deep learning models is evaluated using accuracy, precision, recall, and the F1 score. The LSTM model achieves the highest accuracy of 87.33%, while the CNN model achieves the lowest accuracy of 68.77%. The GRU and BiLSTM models achieve a 93% recall score for the positive sentiment class. The lowest recall of 44% is obtained by CNN. The CNN model shows poor performance both with the TextBlob and VADER techniques.

Table 7. Results of deep learning models using the VADER technique.

Model	Accuracy	Class	Precision	Recall	F1 Score
CNN	68.77	Positive	77	68	72
		Negative	56	44	50
		Neutral	65	80	72
RNN	82.40	Positive	809	88	89
		Negative	62	66	64
		Neutral	83	82	83
LSTM	87.33	Positive	89	92	90
		Negative	74	75	75
		Neutral	91	87	89
BiLSTM	86.95	Positive	88	93	90
		Negative	76	74	75
		Neutral	91	86	88
GRU	86.48	Positive	88	93	90
		Negative	74	70	72
		Neutral	90	86	88

4.3. Results of Transformer-Based Models

Currently, transformer-based models are very effective and perform well on complex natural language understanding (CNLU) tasks in sentiment classification. Machine learning and deep learning models are also used for sentiment analyses, but machine learning performs well on small datasets and deep learning models require large datasets to achieve a high accuracy.

Table 8 shows the results of transformer-based models using the TextBlob technique. The transformer-based robustly optimized BERT model achieves the lowest accuracy of 93.68% while 96% of recall scores are achieved for positive and neutral classes by RoBERTa. The XLNet model achieves an 85.96% accuracy which is low as compared to the RoBERTa and proposed BERT model. In comparison to any other machine or deep learning model, the proposed approach achieves the highest accuracy of 96.49%. The precision, F1 score, and recall of the proposed approach are also higher than those of others.

The results of transformer-based models are also evaluated using the VADER technique. The proposed approach also performs well using the VADER technique with the highest accuracy, as shown in Table 9. The proposed approach understands full contextual content, gives importance to relevant parts of textual data, and makes efficient predictions. The RoBERTa and XLNet transformer-based models achieve 59.59% and 68.51% accuracy scores, respectively. Using the VADER technique, the proposed method achieved a 93.37% accuracy which is higher than all of the other transformer-based models when used with VADER. The other performance metrics, such as precision, recall, and the F1 score, achieved by the proposed model are also better than the other models.

Table 8. Performance of transformer-based models using the TextBlob technique.

Model	Accuracy	Class	Precision	Recall	F1 Score
RoBERTa	93.68	Positive	95	96	93
		Negative	84	85	85
		Neutral	95	96	96
XLNet	85.96	Positive	93	83	87
		Negative	66	77	71
		Neutral	86	91	89
Proposed BERT	96.49	Positive	96	98	97
		Negative	92	90	91
		Neutral	98	97	98

Table 9. Performance of transformer-based models using the VADER technique.

Model	Accuracy	Class	Precision	Recall	F1 Score
RoBERTa	86.68	Positive	75	79	77
		Negative	88	88	88
		Neutral	90	88	89
XLNet	68.51	Positive	66	72	69
		Negative	25	45	32
		Neutral	85	70	76
Proposed BERT	93.37	Positive	97	92	95
		Negative	87	89	88
		Neutral	93	96	94

Table 10 shows the correct and wrong predictions by deep learning and BERT models using the TextBlob. Results are given only for the TextBlob technique, as the models perform well using the TextBlob technique. Out of 4000 predictions, the RNN made 3614 correct predictions and 386 wrong predictions. The LSTM made 3718 correct predictions while 282 predictions are wrong. The BiLSTM has 3725 correct and 275 wrong predictions. The GRU shows 3693 correct predictions, compared to 307 wrong ones. Out of 4160 predictions, the XLNet made 3576 correct and 584 wrong predictions. On the other hand, the RoBERTa made 3897 correct and 263 wrong predictions. The BERT made 4015 correct predictions whereas 146 predictions are wrong. The results demonstrate that the BERT model performed better than the machine learning and deep learning models. Only with 2835 correct and 1165 wrong predictions, the only CNN model performed poorly.

Table 10. Correct and wrong predictions by various models using the TextBlob technique.

Model	Correct-Predictions	Wrong-Predictions	Total-Predictions
CNN	2835	1165	4000
RNN	3614	386	4000
LSTM	3718	282	4000
BiLSTM	3725	275	4000
GRU	3693	307	4000

Table 10. *Cont.*

Model	Correct-Predictions	Wrong-Predictions	Total-Predictions
XLNet	3576	584	4160
RoBERTa	3897	263	4160
Proposed BERT	**4015**	**146**	**4161**

4.4. Results of K-Fold Cross-Validation

K-fold cross-validation is the most effective method for assessing the model's robustness and validating its performance. Table 11 shows the results of Transformer-based models with K-fold cross-validation. Experiments show that the proposed BERT model is highly efficient in the sentiment analysis for ChatGPT tweets with an average accuracy of 96.49% using the TextBlob approach with a $\pm$0.01 standard deviation. The proposed model also works well using the VADER approach with a $\pm$0.01 standard deviation. The RoBERTa on the K-fold achieves a 91% accuracy with a $\pm$0.06 standard deviation, while XLNet achieves a 68% accuracy with a $\pm$0.18 standard deviation.

Table 11. K-fold cross-Validation results using TextBlob and VADER approaches.

	Model	Accuracy	Standard Devation
TextBlob	RoBERTa	0.91	$\pm$0.06
	XLNet	0.68	$\pm$0.18
	Proposed BERT	0.95	$\pm$0.01
VADER	RoBERTa	0.85	$\pm$0.02
	XLNet	0.66	$\pm$0.02
	Proposed BERT	0.93	$\pm$0.01

4.5. Topic Modeling Using BERTopic and LDA Method

Topic modeling is an important approach in NLP, as it automatically extracts the most significant topics from textual data. There is a vast amount of unstructured data available on social media, and traditional approaches are incapable of handling such data. Topic modeling can handle and extract meaningful information from unstructured text data efficiently. In Python, topic modeling is applied to the preprocessed data with important libraries to improve the results. Topic modeling is also used to discover related topics from frequently discussed tweets' datasets.

In various NLP tasks, transformer-based models have produced very promising results. BERTopic is a new topic modeling method that employs the BERT transformer model to extract key trends or keywords from large datasets. BERTopic gathers semantic information that better represents topics. BERT extracts contextual and complicated problems more accurately and efficiently. Furthermore, BERTopic extracts relevant recent trends from Twitter. When compared to LDA modeling, LDA is incapable of extracting nuanced and complicated contextual issues from tweets. In comparison to BERTopic, LDA employs outdated techniques and is unable to extract current patterns. However, BERTopic is a better choice for topic modeling for large datasets.

LDA [74] is an approach used for topic modeling in NLP problems. It is easy to use, efficient, and faster than other approaches for topic modeling. LDA modeling is performed on textual data, and then a document term matrix is created that shows the frequency of each term in a document. The BoW features are utilized to understand the most crucial terms in a document. After that, the most prominent keywords are extracted from ChatGPT tweets using BERTopic, and the LDA are shown in Figure 4.

(**a**) (**b**)

Figure 4. Comparison of LDA-based and BERT-based topic modeling techniques through word clouds: (**a**) Visualization of tweets using LDA topic modeling, and (**b**) Visualization of tweets using BERTopic modeling.

Figure 5 depicts the most prominent topics extracted by BERTopic. First, we load the BERT model and associated tokenizers. The tweet data are then preprocessed to extract the embeddings for the BERT model. Then, for dimension reduction or clustering, we used k-means clustering and the principal component analysis (PCA). The BERT model was used to extract the most prominent topics, which were then displayed in a scatter plot.

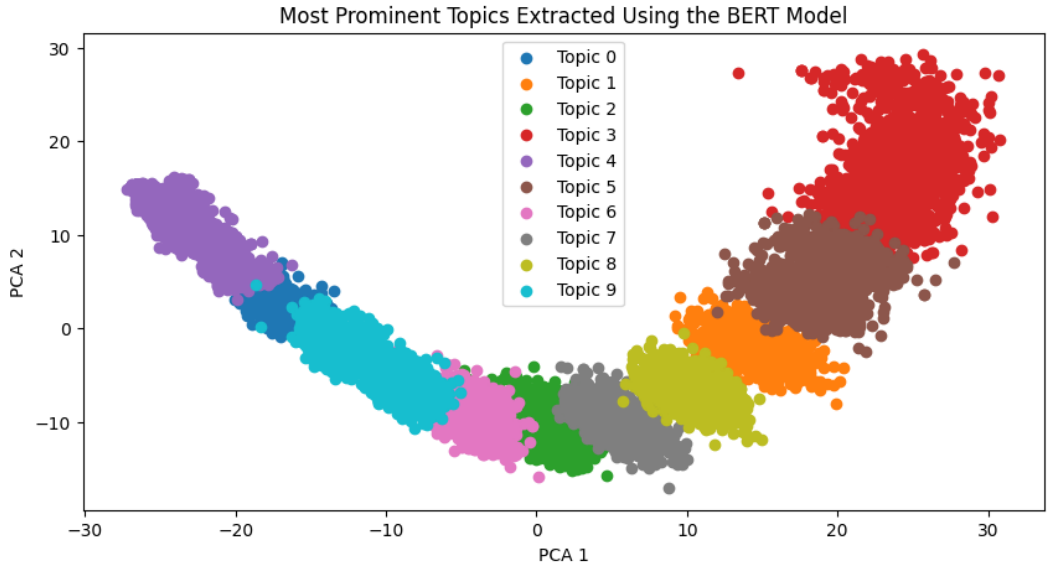

Figure 5. Most Prominent Topics extracted from ChatGPT Tweets using BERTopic.

Figure 6 expresses the content or words of the top ten positive and negative topics and their frequency. The word ChatGPT is mostly discussed in the Twitter tweets in a positive context, and negative words like fake and wrong are discussed but less. The words good, best, and love have the lowest frequency in the top ten topics.

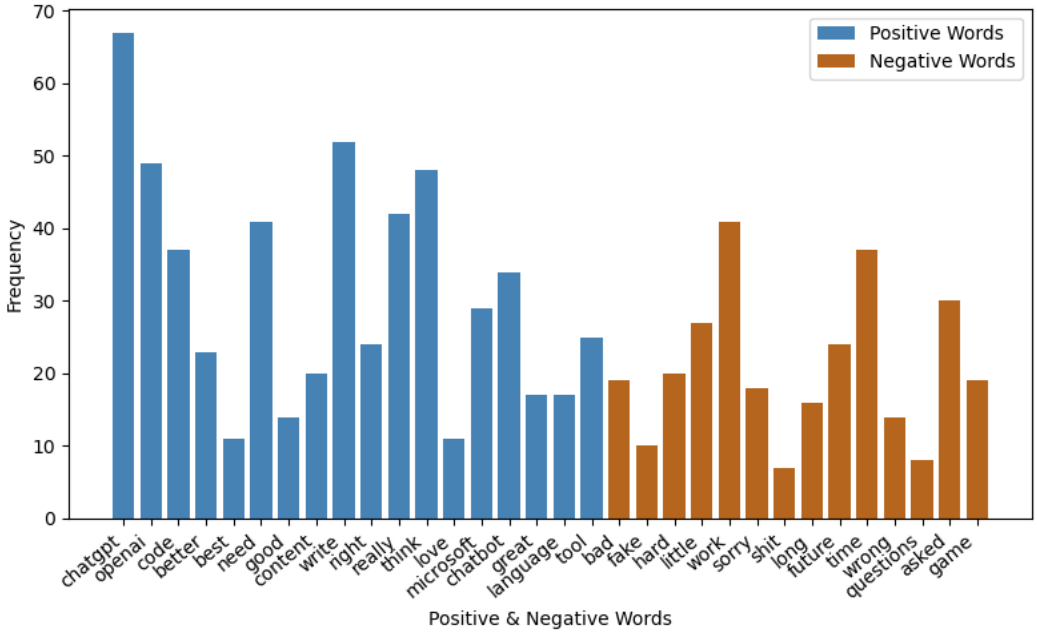

Figure 6. Words extracted from top ten topics with their frequency using the LDA model.

Figures 7 and 8 show the most discussed positive and negative topics, extracted from the ChatGPT tweets using the LDA approach with BoW features. These Figures illustrate positive and negative words in the context of various topics. The users shared their opinions regarding ChatGPT on social media platforms like Twitter. The user posted positive or negative opinions about ChatGPT. The authors extract these tweets from Twitter and perform an analysis to analyze how people feel about or discuss this technology. The authors used LDA-based Topic modeling to extract the most prominent keywords from the tweets. These keywords provide important themes to understand the main context and identify the emotions; they also capture semantic meanings. In the tweets, the word "good" indicates a cheerful mood. It represents anything beneficial or pleasurable. The majority of the time, "good" refers to a positive quality. It is classified as positive sentiment in the sentiment analysis because this inference is generally understood to be positive. It is important to clarify that these words are not inherently positive or negative; rather, their categorization depends on the positive or negative topics they are associated with. For instance, words like "better", "best", and "good" are included in positive topics and are used in a positive context within GPT. Better indicates an advance over a previous state or condition, indicating a positive development. ChatGPT is frequently spoken of favorably due to its features and potential applications in a variety of industries. The development of AI language models like ChatGPT is demonstrated by their ability to comprehend and generate text responses that resemble human responses. ChatGPT allows users to partake in entertaining and engaging conversations. On the other hand, ChatGPT in the negative context indicates that it sometimes produces irrelevant or incorrect results, raises privacy concerns, and an excessive dependence on ChatGPT may impair the ability to think critically and solve technical problems. Social media users frequently use words like "bad", "wrong", "little", and "hot" in a negative sense, aligning with negative topics. Sentiment analysis models can be refined and improved over time based on feedback and real-world data to better capture the nuances of sentiments expressed in different contexts. The performance can be analyzed by policymakers based on these prominent keywords, and they can modify their product according to this.

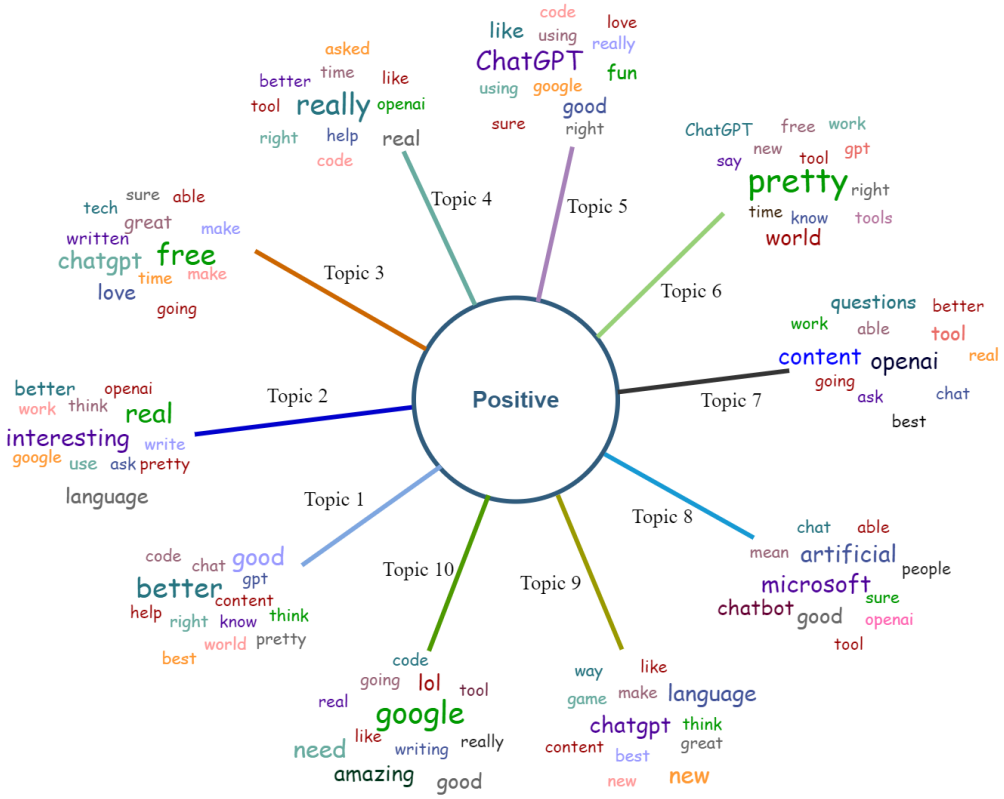

Figure 7. Visualization of highly discussed positive topics.

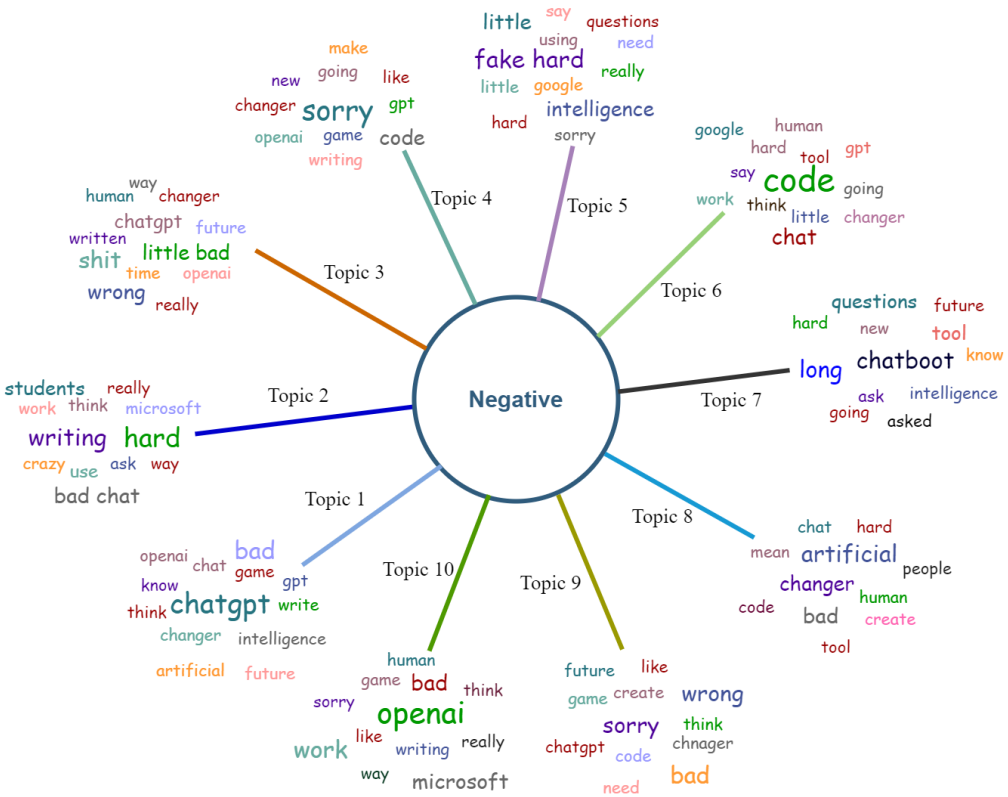

Figure 8. Visualization of highly discussed negative topics.

4.6. Comparison of Proposed Approach with Machine Learning Models Using Statistical Test

The comparison between the machine learning and the proposed Transformer-based BERT model is presented in Table 12. Machine learning models are fine-tuned to optimize the results. The authors evaluated the proposed approach using the TexBlob and Vader technique. In all scenarios, the proposed approach rejects the H_o and accepts the H_a, which means that the proposed approach is statistically significant in comparison with other approaches.

Table 12. Statistical test comparison with the proposed model.

	TextBlob			Vader		
Scenario	**Statistics**	***p*-Value**	H_o	**Statistics**	***p*-Value**	H_o
Proposed BERT Vs. SGD	−7.999	0.015	Rejected	−31.128	7.284	Rejected
Proposed BERT Vs. RF	−39.167	3.661	Rejected	−3.695	0.343	Rejected
Proposed BERT Vs. DT	0.633	0.571	Rejected	−34.097	5.545	Rejected
Proposed BERT Vs. ETC	−63.516	8.598	Rejected	−3.43	0.041	Rejected
Proposed BERT Vs. KNN	−8.225	0.003	Rejected	−6.140	0.008	Rejected
Proposed BERT Vs. SVM	−9.792	0.002	Rejected	−3.257	0.047	Rejected
Proposed BERT Vs. GBM	−9.845	0.002	Rejected	−3.313	0.045	Rejected
Proposed BERT Vs. LR	−17.691	0.000	Rejected	−3.368	0.043	Rejected

4.7. Performance Comparison with State-of-the-Art Studies

For evaluating the robustness and efficiency of the proposed approach, its performance is compared with the state-of-the-art existing studies. Table 13 shows the results of state-of-the-art studies. The study [26] used machine learning models for a sentiment analysis and LR performed well with 83% accuracy. Khalid et al. [27] performed an analysis on Twitter data using an ensemble of machine learning models and achieved 93% accuracy with the BBSVM model. Another study [75] carried out a sentiment analysis on Twitter data using machine learning models. Machine learning models do not perform well due to small datasets and show poor accuracy. As a result, the authors used transformer-based models for the sentiment analysis. For example, Bello et al. [33] used the BERT model on tweets. The proposed BERT model utilizes contextual information to produce a vector representation. When integrated with neural network classifiers such as CNN, RNN, or BiLSTM for prediction, it attains an accuracy rate of 93% and an F measure of 95%. The BiLSTM model exhibits some shortcomings, one of which is its inability to effectively capture the underlying contextual nuances of individual words. Other authors, such as [34,35], used the BERT models for the sentiment analysis with various datasets. They conducted an evaluation of the efficacy of Google's BERT method in comparison to other machine learning methods. Moreover, this study investigates the Bert architecture, which received pre-training on two natural language processing tasks, namely Masked language Modeling and sentence Prediction. The Random Forest (RF) is commonly employed as a benchmark for evaluating the performance of the BERT language model due to its superior performance among various machine learning methods. Previous methodologies are mostly on natural language techniques for the classification and analysis of tweets, yielding insufficient results. The aforementioned prior research indicates the need for an approach that can effectively analyze tweets based on their precise classification. The performance analysis indicates that the proposed BERT model shows efficient results with a 96.49% accuracy and outperforms existing studies.

Table 13. Comparison of proposed approach with state-of-the-art existing studies.

Authors	Model	Dataset	Accuracy	Publication
Rustam et al. [26]	Logistic Regression	App reviews	83%	2020
Khalid et al. [27]	GBSVM	Twitter Data	93%	2020
Wadhwa et al. [75]	Logistic Regression	Twitter Data	86.51%	2021
Bello et al. [33]	BERT	Twitter Data	93%	2022
Catelli et al. [34]	BERT	E-commerce reviews	75%	2021
Patel et al. [35]	BERT	Reviews	83	2022
Proposed	**BERT**	**Twitter Data**	**96.49%**	**2023**

4.8. Validation of Proposed Approach on Additional Dataset

The validation of the proposed approach is carried out using an additional public benchmark dataset. For this purpose, experiments are performed on the well-known SemEval2013 dataset [76]. The proposed TextBlob+BERT approach is applied to the SemEvel2013 dataset, where TextBlob generates new labels for the dataset, and the proposed BERT model performs classification. Moreover, experiments are also done using the original labels of SemEvel2013. Experimental results are presented in Table 14 which indicate the superior performance of the proposed approach. It can be observed that the proposed approach performs significantly well on the SemEvel2013 dataset with a 0.97 accuracy score when labels are assigned using the TextBlob and BERT is used for classification. For the second set of experiments which involves using the original labels of the SemEvel2013 dataset, LR shows the best performance with a 0.65 accuracy score.

Table 14. Experimental results on the SemEvel2013 dataset.

Approach	Accuracy	Class	Precision	Recall	F1 Score
TextBlob + BERT	0.97	Negative	0.97	0.91	0.94
		Neutral	0.98	0.99	0.98
		Positive	0.96	0.98	0.97
		macro avg	0.97	0.96	0.97
		weighted avg	0.97	0.97	0.97
Original + LR	0.65	Negative	0.65	0.47	0.54
		Neutral	0.63	0.72	0.67
		Positive	0.69	0.65	0.67
		macro avg	0.65	0.62	0.63
		weighted avg	0.65	0.65	0.65

4.9. Statistical Significance Test

This study performs a statistical significance *t*-Test to show the significance of the proposed approach. For the statistical test, several scenarios are considered, as mentioned in Table 15. The *t*-test shows the significance of one approach on the other by accepting or rejecting the null hypothesis (H_o). In this study, we consider two cases [77]:

- Null Hypothesis $(H_o) => \mu 1 = \mu 2$: The population means of the proposed approach's results is equal to the compared approach's results. (No statistical significance)
- Alternative Hypothesis $(H_a) => \mu 1 \neq \mu 2$: The population means of the proposed approach's results is not equal to the compared approach's results. (Proposal approach is statistically significant)

Table 15. Statistical significance *t*-test.

Scenario	Statistic	*p*-Value	H_o
Proposed BERT Vs. RoBERTa	3.304	3.304	Rejected
Proposed BERT Vs. XLNet	7.292	0.0003	Rejected
Proposed BERT Vs. GRU	4.481	0.004	Rejected
Proposed BERT Vs. BiLSTM	2.621	0.003	Rejected
Proposed BERT Vs. LSTM	2.510	0.045	Rejected
Proposed BERT Vs. RNN	6.474	0.000	Rejected
Proposed BERT Vs. CNN	8.980	0.000	Rejected

The *t*-test can be interpreted as if the output *p*-value is greater than the alpha value (0.05), it indicates that the H_o is accepted and there is no statistical significance. Moreover, if the *p*-value is less than the alpha value, it indicates that H_o is rejected and H_a is accepted which means that there is statistical significance between the compared results. We perform a *t*-test on results using Textblob and compare all models' performances. In all scenarios, the proposed approach rejects the H_o and accepted the H_a, which means that the proposed approach is statistically significant in comparison with other approaches.

4.10. Discussion

In this study, we observed that the majority of sentiment towards chatGPT was positive, indicating a generally favorable perception of the tool. This aligns with the notion that chatGPT has gained significant attention and popularity on various online platforms. The positive sentiment towards chatGPT can be attributed to its advanced language generation capabilities and its ability to engage in human-like conversations. Figure 9 shows the sentiment ratio for chatGPT.

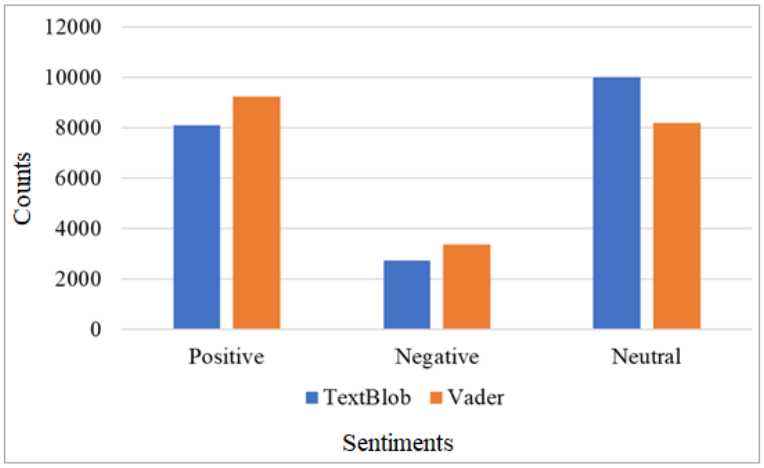

Figure 9. Sentiment ratio in extracted data.

The positive sentiment towards chatGPT is also reflected in the widespread discussions and positive experiences shared by individuals, communities, and social media platforms. People are fascinated by its ability to understand and respond effectively, enhancing user engagement and satisfaction. However, it is important to acknowledge that there are varying opinions and discussions surrounding chatGPT. While most sentiments are positive, some individuals criticize its services and express negative sentiments, particularly concerning its suitability for students. These discussions highlight the need for a further analysis and exploration to address any concerns and improve the tool's effectiveness.

If students rely excessively on ChatGPT, they will lose their capacity to independently compose or generate answers to questions. Students' writing skills may not have improved if they used ChatGPT for projects. As the exam date approaches, individuals have difficulty writing and responding to queries efficiently. There is also the possibility of receiving erroneous information, becoming excessively reliant on technology, and having poor reasoning skills when utilizing ChatGPT. When utilized for personalized learning, ChatGPT may necessitate a comprehensive understanding of the course being taken, the learning preferences of each individual student, and the cultural context in which the students are based. Another negative sentiment regarding ChatGPT is that when students completely rely on AI chatbots to search for specific information about their subject, their level of knowledge does not improve. They cannot advance or increase the topic's knowledge, and it is extremely difficult to maintain concentration when studying. Additionally, students enter data into ChatGPT while looking up specific queries, which could pose a security concern because ChatGPT stores the data that users submit. Over fifty percent of students are motivated to cheat and use ChatGPT to generate information for their submissions. While most students did not admit to using ChatGPT in their writing, integrity may be compromised when ChatGPT generates text.

Additionally, we conducted an analysis using an external sentiment analysis tool called SentimentViz [78]. This tool allowed us to visualize people's perceptions of ChatGPT based on their data. The sentiment analysis results obtained from SentimentViz complemented and validated the findings of the proposed approach. Figure 10 presents visual representations of the sentiment expressed by individuals regarding ChatGPT. This visualization provides further support for the positive sentiment observed in our study and reinforces the credibility of our results.

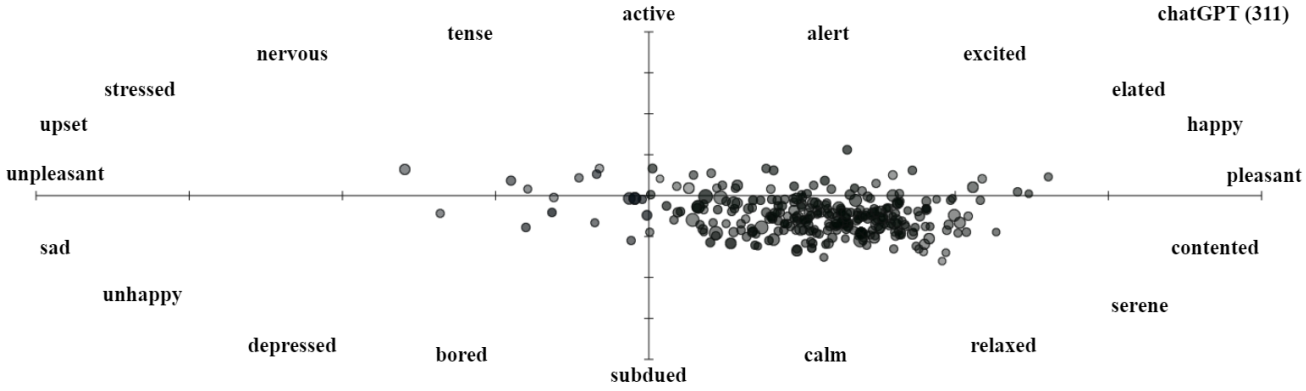

Figure 10. SentimentViz output for chatGPT sentiment.

Discussions regarding the set RQs for this study are also given here.

i. **RQ1:** What are people's sentiments about ChatGPT technology?
Response: The authors analyzed a large dataset of tweets and were able to determine how individuals feel about ChatGPT technology. The results indicate that users have mixed feelings about ChatGPT, with some expressing positive opinions and others expressing negative views. These results provide useful information about how the public perceives ChatGPT and can assist researchers and developers in understanding the chatbot's strengths and weaknesses. The favorable perception of chatGPT is attributable to its advanced language generation features and its ability to become involved in human-like interactions. Individuals are attracted by its cognitive power as well as its ability to effectively respond, thereby increasing user interest and satisfaction. The positive sentiments, like the new openai ChatGPT, writes user-generated content in a better way; it is a great language tool that codes you for your specific queries, etc.

ii. **RQ2:** Which classification model is most effective, such as the proposed transformer-based models, machine learning-based models, and deep learning-based models, for analyzing sentiments about ChatGPT tweets?
Response: The experiments indicate that transformer-based BERT models are more effective and accurate for analyzing sentiments about the ChatGPT tweets. Since transformers make use of self-attention mechanisms, they give the same amount of attention to each component of the sequence that they are processing. They have the ability to virtually process any kind of sequential information. When it comes to natural language processing (NLP), the BERT model takes into account the context of words in both directions (left to right and right to left). Transformers have an in-depth understanding of the meanings of words and are useful for complex problems. In contrast, manual feature engineering, rigorous preprocessing, and a limited dataset are required for machine learning in order to improve accuracy. Additionally, deep learning has a less accurate automatic feature extraction method.

iii. **RQ3:** What are the impacts of ChatGPT on student learning?
Response: The findings show that ChatGPT may have a significant impact on students' learning. ChatGPT's learning capabilities can help students learn when they do not attend school. ChatGPT is not recommended to be used as a substitute for analytical thinking and creative work, but also as a tool to develop research and writing skills. Students' writing skills may not have improved if they relied completely on ChatGPT. There is also the possibility of receiving erroneous information, becoming excessively reliant on technology, and having poor reasoning skills.

iv. **RQ4:** What role does topic modeling play in the sentiment analysis of social media tweets?
Response: Topic modeling refers to an unsupervised statistical method to assess whether or not a particular batch of documents contains any "topics" that are more generic in nature. In order to create a summary that is the most accurate depiction of the document's contents, it extracts the text for commonly used words and phrases. There is a vast amount of unstructured data related to OpenAI ChatGPT, and traditional approaches are incapable of handling such data. Topic modeling can handle and extract meaningful information from unstructured text data efficiently. LDA-based modeling extracts the most discussed topics and prominent positive or negative keywords. It also provides clear information from the large corpus, which is very time-consuming if an individual extracts topics manually.

5. Conclusions

This study conducted a sentiment analysis on ChatGPT-related tweets to gain insight into people's perceptions and opinions. By analyzing a large dataset of tweets, we were able to identify the overall sentiment expressed by users towards ChatGPT. The findings indicate that there are mixed sentiments among users, with some expressing positive views and others expressing negative views about ChatGPT. These results provide valuable insights into the public perception of ChatGPT and can help researchers and developers understand the strengths and weaknesses of the chatbot. Further, this study utilized the BERT model to analyze tweets related to ChatGPT. The BERT model proved to be effective in understanding and classifying sentiments expressed in these tweets. By employing the BERT model, we were able to accurately classify sentiments and gain a deeper understanding of the overall sentiment trends surrounding ChatGPT.

The experimental results demonstrate the outstanding performance of the proposed model, achieving an accuracy of 94.96%. This performance is further validated through k-fold cross-validation and comparison with existing state-of-the-art studies. Our conclusions indicate that the majority of people expressed positive sentiments towards the ChatGPT tool, while a minority had negative sentiments. It was observed that many users appreciate the tool for its assistance across various domains. However, some individuals criticized

the ChatGPT tool's services, particularly its suitability for students, expressing negative sentiments in this regard.

This study recognizes the limitation of a relatively small dataset, comprising only 21,515 tweets, which may restrict comprehensive insights. To overcome this limitation, future research will prioritize the collection of a larger volume of data from Twitter and other social media platforms to gain a more accurate understanding of people's perceptions of the trending chatGPT tool. Moreover, the study aims to develop a machine learning approach that incorporates the sentiment analysis, enabling exploration of how such technologies can be developed to mitigate potential societal harm and ensure responsible deployment.

Author Contributions: Conceptualization, S.R. and M.M.; Data curation, M.M. and F.R.; Formal analysis, S.R., F.R., R.S. and I.d.l.T.D.; Funding acquisition, I.d.l.T.D.; Investigation, V.C. and M.G.V.; Methodology, F.R., M.M. and R.S.; Project administration, R.S. and V.C.; Resources, M.G.V. and J.B.B.; Software, M.G.V. and J.B.B.; Supervision, I.d.l.T.D. and I.A.; Validation, J.B.B. and I.A.; Visualization, R.S. and V.C.; Writing—original draft, M.M., R.S., F.R. and S.R.; Writing—review & editing, I.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European University of Atlantic.

Data Availability Statement: https://www.kaggle.com/datasets/furqanrustam118/chatgpt-tweets.

Conflicts of Interest: The authors declare no conflict of interests.

References

1. Meshram, S.; Naik, N.; Megha, V.; More, T.; Kharche, S. Conversational AI: Chatbots. In Proceedings of the 2021 International Conference on Intelligent Technologies (CONIT), Hubli, India, 25–27 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6.
2. The Future of Chatbots: 10 Trends, Latest Stats & Market Size. Available online: https://onix-systems.com/blog/6-chatbot-trends-that-are-bringing-the-future-closer (accessed on 23 May 2023).
3. Size of the Chatbot Market Worldwide from 2021 to 2030. Available online: https://www.statista.com/statistics/656596/worldwide-chatbot-market/ (accessed on 23 May 2023).
4. Chatbot Market in 2022: Stats, Trends, and Companies in the Growing AI Chatbot Industry. Available online: https://www.insiderintelligence.com/insights/chatbot-market-stats-trends/ (accessed on 23 May 2023).
5. Malinka, K.; Perešíni, M.; Firc, A.; Hujňák, O.; Januš, F. On the educational impact of ChatGPT: Is Artificial Intelligence ready to obtain a university degree? *arXiv* **2023**, arXiv:2303.11146.
6. George, A.S.; George, A.H. A review of ChatGPT AI's impact on several business sectors. *Partners Univers. Int. Innov. J.* **2023**, *1*, 9–23.
7. Lund, B.D.; Wang, T.; Mannuru, N.R.; Nie, B.; Shimray, S.; Wang, Z. ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *J. Assoc. Inf. Sci. Technol.* **2023**, *74*, 570–581.
8. Kirmani, A.R. Artificial Intelligence-Enabled Science Poetry. *ACS Energy Lett.* **2022**, *8*, 574–576.
9. Cotton, D.R.; Cotton, P.A.; Shipway, J.R. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innov. Educ. Teach. Int.* **2023**, 1–12. [CrossRef]
10. Tlili, A.; Shehata, B.; Adarkwah, M.A.; Bozkurt, A.; Hickey, D.T.; Huang, R.; Agyemang, B. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learn. Environ.* **2023**, *10*, 15.
11. Edtech Chegg Tumbles as ChatGPT Threat Prompts Revenue Warning. Available online: https://www.reuters.com/markets/us/edtech-chegg-slumps-revenue-warning-chatgpt-threatens-growth-2023-05-02/ (accessed on 23 May 2023).
12. Liu, B. *Sentiment Analysis and Opinion Mining*; Synthesis Lectures on Human Language Technologies; Springer: Cham, Switzerland, 2012; Volume 5, 167p.
13. Medhat, W.; Hassan, A.; Korashy, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Eng. J.* **2014**, *5*, 1093–1113.
14. Hussein, D.M.E.D.M. A survey on sentiment analysis challenges. *J. King Saud Univ.-Eng. Sci.* **2018**, *30*, 330–338.
15. Lee, E.; Rustam, F.; Ashraf, I.; Washington, P.B.; Narra, M.; Shafique, R. Inquest of Current Situation in Afghanistan Under Taliban Rule Using Sentiment Analysis and Volume Analysis. *IEEE Access* **2022**, *10*, 10333–10348.
16. Lee, E.; Rustam, F.; Washington, P.B.; El Barakaz, F.; Aljedaani, W.; Ashraf, I. Racism detection by analyzing differential opinions through sentiment analysis of tweets using stacked ensemble gcr-nn model. *IEEE Access* **2022**, *10*, 9717–9728. [CrossRef]
17. Mujahid, M.; Lee, E.; Rustam, F.; Washington, P.B.; Ullah, S.; Reshi, A.A.; Ashraf, I. Sentiment analysis and topic modeling on tweets about online education during COVID-19. *Appl. Sci.* **2021**, *11*, 8438. [CrossRef]
18. Tran, A.D.; Pallant, J.I.; Johnson, L.W. Exploring the impact of chatbots on consumer sentiment and expectations in retail. *J. Retail. Consum. Serv.* **2021**, *63*, 102718. [CrossRef]

19. Muneshwara, M.; Swetha, M.; Rohidekar, M.P.; AB, M.P. Implementation of Therapy Bot for Potential Users With Depression During Covid-19 Using Sentiment Analysis. *J. Posit. Sch. Psychol.* **2022**, *6*, 7816–7826.
20. Parimala, M.; Swarna Priya, R.; Praveen Kumar Reddy, M.; Lal Chowdhary, C.; Kumar Poluru, R.; Khan, S. Spatiotemporal-based sentiment analysis on tweets for risk assessment of event using deep learning approach. *Softw. Pract. Exp.* **2021**, *51*, 550–570. [CrossRef]
21. Aslam, N.; Rustam, F.; Lee, E.; Washington, P.B.; Ashraf, I. Sentiment analysis and emotion detection on cryptocurrency related Tweets using ensemble LSTM-GRU Model. *IEEE Access* **2022**, *10*, 39313–39324. [CrossRef]
22. Aslam, N.; Xia, K.; Rustam, F.; Lee, E.; Ashraf, I. Self voting classification model for online meeting app review sentiment analysis and topic modeling. *PeerJ Comput. Sci.* **2022**, *8*, e1141. [CrossRef] [PubMed]
23. Araujo, A.F.; Gôlo, M.P.; Marcacini, R.M. Opinion mining for app reviews: An analysis of textual representation and predictive models. *Autom. Softw. Eng.* **2022**, *29*, 1–30. [CrossRef]
24. Aljedaani, W.; Mkaouer, M.W.; Ludi, S.; Javed, Y. Automatic classification of accessibility user reviews in android apps. In Proceedings of the 2022 7th international conference on data science and machine learning applications (CDMA), Riyadh, Saudi Arabia, 1–3 March 2022; IEEE: Piscataway, NJ, USA, 2022, pp. 133–138.
25. Naeem, M.Z.; Rustam, F.; Mehmood, A.; Ashraf, I.; Choi, G.S. Classification of movie reviews using term frequency-inverse document frequency and optimized machine learning algorithms. *PeerJ Comput. Sci.* **2022**, *8*, e914. [CrossRef]
26. Rustam, F.; Mehmood, A.; Ahmad, M.; Ullah, S.; Khan, D.M.; Choi, G.S. Classification of shopify app user reviews using novel multi text features. *IEEE Access* **2020**, *8*, 30234–30244. [CrossRef]
27. Khalid, M.; Ashraf, I.; Mehmood, A.; Ullah, S.; Ahmad, M.; Choi, G.S. GBSVM: Sentiment classification from unstructured reviews using ensemble classifier. *Appl. Sci.* **2020**, *10*, 2788. [CrossRef]
28. Umer, M.; Ashraf, I.; Mehmood, A.; Ullah, S.; Choi, G.S. Predicting numeric ratings for google apps using text features and ensemble learning. *ETRI J.* **2021**, *43*, 95–108. [CrossRef]
29. Rehan, M.S.; Rustam, F.; Ullah, S.; Hussain, S.; Mehmood, A.; Choi, G.S. Employees reviews classification and evaluation (ERCE) model using supervised machine learning approaches. *J. Ambient Intell. Humaniz. Comput.* **2022**, *13*, 3119–3136. [CrossRef]
30. Al Kilani, N.; Tailakh, R.; Hanani, A. Automatic classification of apps reviews for requirement engineering: Exploring the customers need from healthcare applications. In Proceedings of the 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, 22–25 October 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 541–548.
31. Srisopha, K.; Phonsom, C.; Lin, K.; Boehm, B. Same app, different countries: A preliminary user reviews study on most downloaded ios apps. In Proceedings of the 2019 IEEE International Conference on Software Maintenance and Evolution (ICSME), Cleveland, OH, USA, 29 September–4 October 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 76–80.
32. Hossain, M.S.; Rahman, M.F. Sentiment analysis and review rating prediction of the users of Bangladeshi Shopping Apps. In *Developing Relationships, Personalization, and Data Herald in Marketing 5.0*; IGI Global: Pennsylvania, PA USA, 2022; pp. 33–56.
33. Bello, A.; Ng, S.C.; Leung, M.F. A BERT Framework to Sentiment Analysis of Tweets. *Sensors* **2023**, *23*, 506. [CrossRef]
34. Catelli, R.; Pelosi, S.; Esposito, M. Lexicon-based vs. Bert-based sentiment analysis: A comparative study in Italian. *Electronics* **2022**, *11*, 374. [CrossRef]
35. Patel, A.; Oza, P.; Agrawal, S. Sentiment Analysis of Customer Feedback and Reviews for Airline Services using Language Representation Model. *Procedia Comput. Sci.* **2023**, *218*, 2459–2467. [CrossRef]
36. Mujahid, M.; Kanwal, K.; Rustam, F.; Aljadani, W.; Ashraf, I. Arabic ChatGPT Tweets Classification using RoBERTa and BERT Ensemble Model. *Acm Trans. Asian-Low-Resour. Lang. Inf. Process.* **2023**. [CrossRef]
37. Bonifazi, G.; Cauteruccio, F.; Corradini, E.; Marchetti, M.; Sciarretta, L.; Ursino, D.; Virgili, L. A Space-Time Framework for Sentiment Scope Analysis in Social Media. *Big Data Cogn. Comput.* **2022**, *6*, 130. [CrossRef]
38. Bonifazi, G.; Corradini, E.; Ursino, D.; Virgili, L. Modeling, Evaluating, and Applying the eWoM Power of Reddit Posts. *Big Data Cogn. Comput.* **2023**, *7*, 47. [CrossRef]
39. Messaoud, M.B.; Jenhani, I.; Jemaa, N.B.; Mkaouer, M.W. A multi-label active learning approach for mobile app user review classification. In Proceedings of the Knowledge Science, Engineering and Management: 12th International Conference, KSEM 2019, Athens, Greece, 28–30 August 2019; Proceedings, Part I 12; Springer: Berlin/Heidelberg, Germany, 2019; pp. 805–816.
40. Fuad, A.; Al-Yahya, M. Analysis and classification of mobile apps using topic modeling: A case study on Google Play Arabic apps. *Complexity* **2021**, *2021*, 1–12. [CrossRef]
41. Venkatakrishnan, S.; Kaushik, A.; Verma, J.K. Sentiment analysis on google play store data using deep learning. In *Applications of Machine Learning*; Springer: Singapore, 2020; pp. 15–30.
42. Alam, S.; Yao, N. The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. *Comput. Math. Organ. Theory* **2019**, *25*, 319–335. [CrossRef]
43. Vijayarani, S.; Ilamathi, M.J.; Nithya, M. Preprocessing techniques for text mining-an overview. *Int. J. Comput. Sci. Commun. Netw.* **2015**, *5*, 7–16.
44. R, S.; Mujahid, M.; Rustam, F.; Mallampati, B.; Chunduri, V.; de la Torre Díez, I.; Ashraf, I. Bidirectional encoder representations from transformers and deep learning model for analyzing smartphone-related tweets. *PeerJ Comput. Sci.* **2023**, *9*, e1432. [CrossRef]
45. Kadhim, A.I. An evaluation of preprocessing techniques for text classification. *Int. J. Comput. Sci. Inf. Secur.* **2018**, *16*, 22–32.
46. Loria, S. Textblob Documentation. Release 0.15. 2018. Volume 2. Available online: https://buildmedia.readthedocs.org/media/pdf/textblob/latest/textblob.pdf (accessed on 23 May 2023).

47. Borg, A.; Boldt, M. Using VADER sentiment and SVM for predicting customer response sentiment. *Expert Syst. Appl.* **2020**, *162*, 113746. [CrossRef]
48. Karamibekr, M.; Ghorbani, A.A. Sentiment analysis of social issues. In Proceedings of the 2012 International Conference on Social Informatics, Alexandria, VA, USA, 14–16 December 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 215–221.
49. Kumar, Y.; Koul, A.; Singla, R.; Ijaz, M.F. Artificial intelligence in disease diagnosis: A systematic literature review, synthesizing framework and future research agenda. *J. Ambient. Intell. Humaniz. Comput.* **2023**, *14*, 8459–8486 [CrossRef]
50. Shafique, R.; Aljedaani, W.; Rustam, F.; Lee, E.; Mehmood, A.; Choi, G.S. Role of Artificial Intelligence in Online Education: A Systematic Mapping Study. *IEEE Access* **2023**, *11*, 52570–52584. [CrossRef]
51. George, A.; Ravindran, A.; Mendieta, M.; Tabkhi, H. Mez: An adaptive messaging system for latency-sensitive multi-camera machine vision at the iot edge. *IEEE Access* **2021**, *9*, 21457–21473. [CrossRef]
52. Ravindran, A.; George, A. An edge datastore architecture for Latency-Critical distributed machine vision applications. In Proceedings of the USENIX Workshop on Hot Topics in Edge Computing (HotEdge 18), Boston, MA, USA, 10 July 2018.
53. Kadhim, A.I. Survey on supervised machine learning techniques for automatic text classification. *Artif. Intell. Rev.* **2019**, *52*, 273–292. [CrossRef]
54. Chen, H.; Wu, L.; Chen, J.; Lu, W.; Ding, J. A comparative study of automated legal text classification using random forests and deep learning. *Inf. Process. Manag.* **2022**, *59*, 102798. [CrossRef]
55. Schröder, C.; Niekler, A. A survey of active learning for text classification using deep neural networks. *arXiv* **2020**, arXiv:2008.07267.
56. Prabhat, A.; Khullar, V. Sentiment classification on big data using Naïve Bayes and logistic regression. In Proceedings of the 2017 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 5–7 January 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–5.
57. Valencia, F.; Gómez-Espinosa, A.; Valdés-Aguirre, B. Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. *Entropy* **2019**, *21*, 589. [CrossRef] [PubMed]
58. Zharmagambetov, A.S.; Pak, A.A. Sentiment analysis of a document using deep learning approach and decision trees. In Proceedings of the 2015 Twelve International Conference on Electronics Computer and Computation (ICECCO), Almaty, Kazakhstan, 27–30 September 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1–4.
59. Shah, K.; Patel, H.; Sanghvi, D.; Shah, M. A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augment. Hum. Res.* **2020**, *5*, 12. [CrossRef]
60. Tiwari, D.; Singh, N. Ensemble approach for twitter sentiment analysis. *IJ Inf. Technol. Comput. Sci.* **2019**, *8*, 20–26. [CrossRef]
61. Arya, V.; Mishra, A.K.M.; González-Briones, A. Analysis of sentiments on the onset of COVID-19 using machine learning techniques. *ADCAIJ Adv. Distrib. Comput. Artif. Intell. J.* **2022**, *11*, 45–63. [CrossRef]
62. Severyn, A.; Moschitti, A. Unitn: Training deep convolutional neural network for twitter sentiment classification. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, CO, USA, 4–5 June 2015; pp. 464–469.
63. Seo, S.; Kim, C.; Kim, H.; Mo, K.; Kang, P. Comparative study of deep learning-based sentiment classification. *IEEE Access* **2020**, *8*, 6861–6875. [CrossRef]
64. Nowak, J.; Taspinar, A.; Scherer, R. LSTM recurrent neural networks for short text and sentiment classification. In Proceedings of the Artificial Intelligence and Soft Computing: 16th International Conference, ICAISC 2017, Zakopane, Poland, 11–15 June 2017; Proceedings, Part II 16; Springer: Cham, Switzerland, 2017; pp. 553–562.
65. Mujahid, M.; Rustam, F.; Alasim, F.; Siddique, M.; Ashraf, I. What people think about fast food: Opinions analysis and LDA modeling on fast food restaurants using unstructured tweets. *PeerJ Comput. Sci.* **2023**, *9*, e1193. [CrossRef]
66. Tang, D.; Qin, B.; Liu, T. Document modeling with gated recurrent neural network for sentiment classification. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1422–1432.
67. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
68. Tenney, I.; Das, D.; Pavlick, E. BERT rediscovers the classical NLP pipeline. *arXiv* **2019**, arXiv:1905.05950.
69. González-Carvajal, S.; Garrido-Merchán, E.C. Comparing BERT against traditional machine learning text classification. *arXiv* **2020**, arXiv:2005.13012.
70. Cruz, J.C.B.; Cheng, C. Establishing baselines for text classification in low-resource languages. *arXiv* **2020**, arXiv:2005.02068.
71. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 5753–5763.
72. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
73. Amaar, A.; Aljedaani, W.; Rustam, F.; Ullah, S.; Rupapara, V.; Ludi, S. Detection of fake job postings by utilizing machine learning and natural language processing approaches. *Neural Process. Lett.* **2022**, *54*, 2219–2247 [CrossRef]
74. Jelodar, H.; Wang, Y.; Yuan, C.; Feng, X.; Jiang, X.; Li, Y.; Zhao, L. Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimed. Tools Appl.* **2019**, *78*, 15169–15211. [CrossRef]
75. Wadhwa, S.; Babber, K. Performance comparison of classifiers on twitter sentimental analysis. *Eur. J. Eng. Sci. Technol.* **2021**, *4*, 15–24. [CrossRef]

76. SemEvel2013 Dataset. Available online: https://www.kaggle.com/datasets/azzouza2018/semevaldatadets?select=semeval-2013-train-all.csv (accessed on 23 May 2023).
77. Rustam, F.; Ashraf, I.; Mehmood, A.; Ullah, S.; Choi, G.S. Tweets classification on the base of sentiments for US airline companies. *Entropy* **2019**, *21*, 1078. [CrossRef]
78. Sentiment Viz: Tweet Sentiment Visualization. Available online: https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/ (accessed on 23 May 2023).