Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/21/20231118

A Comparison of LSTM and BERT model for sarcasm prediction

Yizhong Ding

SWJTU-Leeds Joint School, Southwest Jiaotong University, Chengdu, Sichuan, 611756, China

sc20yd@leeds.ac.uk

Abstract. Sarcasm prediction is a text analysis task that aims to identify sarcastic and non-sarcastic statements in text. Sarcasm is a figure of speech that uses opposite or contradictory language to express a certain meaning or idea. Sarcasm is usually cryptic, vague, and suggestive, which makes sarcasm prediction a challenging task. In sarcasm prediction projects, techniques of natural language processing are usually leveraged to analyze and classify the text. The main challenge of this task lies in the fact that sarcasm usually has multiple manifestations and needs to consider the contextual and semantic information of the text. The prediction of sarcasm holds significant application value in natural language processing, such as social media analysis, public opinion monitoring, sentiment analysis and so on. In this paper, by controlling variables, the influence of adding the long short-term memory (LSTM) layer and changing the grid structure of the model on the accuracy of prediction results is explored. Moreover, accuracy of the LSTM prediction performance is compared with that of the bidirectional encoder representations from Transformers (BERT) model. At the same time, this paper analyzed and discussed the phenomenon that adding the number of LSTM model layers could not obtain higher prediction accuracy, and the accuracy gap of prediction results between LSTM model and BERT model, and finally obtained relevant conclusions.

Keywords: NLP, sarcasm prediction, LSTM.

1. Introduction

Sarcasm prediction is a significant challenge in natural language processing (NLP), which requires the development of advanced algorithms and models capable of identifying sarcastic language patterns and inferring their intended meaning. Sarcasm is a common expression in everyday communication, especially on social media and online forums. Sarcasm is often expressed by conveying information between the lines, and its meaning is often the opposite of the literal meaning, requiring an understanding of context and context. Therefore, accurately identifying sarcastic expressions in text data is a challenging problem.

In the past few years, interest and research have surged in sarcasm prediction within the domain of the natural language processing. Advanced algorithms and models have been developed to address this challenge, with the aim of detecting and interpreting sarcastic language patterns in various forms of text data. This area of research has gained significance in light of its potential applications in sentiment

DOI: 10.54254/2755-2721/21/20231118

analysis, social media monitoring, and other fields where accurate interpretation of language data is critical. The application of sarcasm prediction involves sentiment analysis, emotion recognition, chatbot development and many other fields [1]. In sentiment analysis, accurately identifying sarcastic expressions in texts can help enterprises to gain insight into customers' real emotions and make adjustments and improvements in time. In emotion recognition, sarcasm prediction can help computers better understand human emotional expressions, thereby improving computer understanding of natural language. In chatbot development, sarcasm prediction can help chatbots better understand the user's intention and emotion, thereby improving the interaction experience.

In addition, sarcasm prediction has practical applications in social media monitoring, customer feedback analysis, online reputation management, and other fields. In social media monitoring, sarcasm prediction can help enterprises better understand consumers' opinions and emotional attitudes toward their brands, and can also help clean up the online environment and identify and avoid phenomena such as cyberbullying in advance [2]. In customer feedback analysis, ironic prediction can help companies better understand the needs and pain points of consumers. In online reputation management, sarcasm prediction can help companies better manage and maintain their brand reputation.

A neural network-based deep learning model is used for sarcasm analysis. Deep models are a branch of machine learning to perform the classification, regression, or generation of data efficiently tasks. Machine learning has its roots in artificial intelligence and computer science [3]. The machine learning field centers around the creation of models and algorithms that enable computers to learn from the data and forecast or decisions autonomously, without the need for explicit programming. This involves the development of mathematical models for extracting underlying patterns and relationships within datasets, to further make accurate predictions or decisions. Machine learning has its origins in the 1950s when researchers first began exploring ways to make machines learn from data. However, it wasn't until the advent of big data and advancements in computing power that machine learning really took off. Machine learning is widely employed in a range of applications, such as image recognition, fraud detection, and referral systems [4]. It provides computers with the capacity to make decisions without explicit programming, revolutionizing various industries such as healthcare, finance, and entertainment. Its popularity can be attributed to its ability to automate complex decision-making processes, improve accuracy, and enable businesses to gain insights from vast amounts of data. As the amount of data generated continues to grow exponentially, machine learning is becoming an increasingly important tool for businesses and researchers alike.

Introduced by Google in 2018, Bidirectional Encoder Representations from Transformers (BERT) is a language model that has gained considerable attention and popularity due to its advanced architecture, which enables it to capture bidirectional contextual relationships between words. Due to its advanced architecture and ability to capture bidirectional contextual relationships between words, it has become widely used and succeeded in various natural language processing tasks. Built on the Transformer architecture, BERT uses a bidirectional training approach to learn context-aware word embeddings, performing at the forefront on various natural language processing tasks. The advantages of BERT lie in its capacity to pre-train on large-scale corpora, allowing for better generalization, and then fine-tune on specific tasks to improve model accuracy [5]. As a variation of recurrent neural network (RNN), by introducing memory cells and gate mechanisms, long short-term memory (LSTM) can effectively capture long-term dependencies between elements in a sequence, averting the vanishing gradient problem in traditional RNNs. The advantages of LSTM lie in its ability to handle sequential data and its potential for deepening the model by stacking multiple LSTM layers, enhancing the model's expressive power [6].

This paper aims to explore the LSTM model, by modifying the number of layers to explore the results prediction accuracy using the LSTM model for sarcasm analysis, and get the best results. At the same time, the BERT model is studied to obtain the sarcasm prediction accuracy of the model. Finally, the results obtained by the LSTM model with the highest accuracy were compared with the prediction

DOI: 10.54254/2755-2721/21/20231118

accuracy results obtained by the BERT model, and the relevant results were analyzed and discussed to draw conclusions.

2. Method

2.1. Dataset

In this project on sarcasm prediction, a JSON file was utilized containing three components: whether the statement is sarcastic or not, headline, and the article link from which the statement was sourced.

The attribute article link was removed for easier analysis of the text. The number of unique words in the dataset is limited to the top 10,000 most frequently occurring words, and the text is tokenized by Tokenizer. Sequences with lengths less than the defined maximum length are filled with <pad> tokens, while longer sequences are truncated. Finally, the distribution of is_sarcastic columns in the dataset is visualized using the countplot function in the Seaborn library, which is used to determine that the categories in the dataset are balanced and can be used for subsequent operations.

2.2. Word to vector

In the LSTM model, pre-trained word embeddings from global vectors for word representation (GloVe) are used. GloVe is a popular method for constructing word embeddings based on co-occurrence statistics in large text corpora. The code loads pre-trained word embeddings from a file named 'glove.6B.200d.txt' and creates a weight matrix for words in the training dataset. For each word in the tokenizer's word index, the corresponding embedding vector is retrieved from the pre-trained embeddings and added to the weight matrix. The matrix is subsequently employed as the initial weights of the LSTM model's embedding layer, allowing the model to preprocess the input data and extract meaningful features from it. In the BERT model, the word-to-vector is based on the BERT tokenizer and pre-trained BERT model weights. The code uses the 'bert-base-uncased' model and tokenizer from the Hugging Face Transformers library. The encoder function takes an input sentence list and returns a list of tokenized sentence IDs using the BERT tokenizer. These tokenized IDs are then used as input to the BERT model's embedding layer, which retrieves the corresponding pre-trained embeddings from the BERT model weights. The benefit of using BERT for word-to-vector encoding is that it has learned on an extensive corpus of text and can provide context-aware embeddings that capture the meaning of words in each sentence or document.

2.3. LSTM, Bi-LSTM, and BERT model

LSTM and Bi-directional LSTM (Bi-LSTM) are popular types of RNN models that have been widely applied in several fields, such as voice recognition and subtitling of images.

LSTM aims at alleviating the vanishing gradient problem in RNNs, which occurs when the gradients propagated through the network become too small to update the weights effectively. The core idea of the LSTM model is to introduce a memory unit that can selectively remember or forget information based on the current input and the previous state. A set of gates regulates the flow of information to update the memory cell: in terms of a cell, the input gate appends information to it and the forget gate deletes information from it, which determines the message to be passed to the next layer.

Bi-LSTM process the input in both directions simultaneously. In contrast to the one direction that the original LSTM only has, to allow the model to capture information about past and future context. This enhanced capability has made Bi-LSTM a popular choice in various natural language processing applications. This feature is achieved by duplicating the LSTM cells and run them forward and back separately. The outputs of the two cells are focused to form the final output. It has been demonstrated that the Bi-LSTM is effective in tasks such as voice recognition and machine translation, where past and future contexts are important [6].

Some of the key features of LSTM and Bi-LSTM include their ability to handle long-term dependencies, their ability to selectively remember or forget information, and their ability to process

DOI: 10.54254/2755-2721/21/20231118

input sequences in both directions. However, they also demand additional computational resources and are harder to train than simpler models such as feedforward neural networks [7].

BERT is a pre-trained language model that makes use of a bidirectional transformer architecture to learn contextualized word representations [5]. The model is pre-trained on a wide corpus of text with a masked language modeling goal, which requires that the pattern predicts hidden words according to their context. One of the key features of BERT is its ability to manage varieties of NLP tasks with the minimal task-specific changes of architecture, achieved through fine-tuning the pre-trained model on task-specific data. Many NLP benchmarks, such as question answering and sentiment analysis, have obtained leading-edge results using BERT.

BERT's transformer architecture uses self-attention to compute contextualized representations of the input sequence, enable the model to process different parts in the sequence and weigh these parts based on their relevance to the current task [8].

Since this paper aims to compare bidirectional LSTM models with BERT models, the model structure of LSTM, that is, the number of layers, will be modified to enhance the model's complexity and expressiveness, help capture more information in the input sequence, and explore the highest accuracy results that bidirectional LSTM models can obtain under different model structures. The results are compared with the BERT model prediction results.

2.4. Evaluation index

In sarcasm analysis prediction, the accuracy, Receiver Operating Characteristic (ROC) curve, and confusion matrix, these three indices will be used for this project.

A widely used metric in evaluating model performance is precision, which measures the percentage of accurate predictions the model has made. To calculate precision, the number of correct predictions is divided by the total number of predictions performed by the model. This metric is particularly useful in applications where accurate prediction of positive instances is critical.

The ROC curve is used to have an evaluation for the binary classification model. The AUC is a widely used metric to measure the model performance, with higher values indicating better performance. This method provides a comprehensive view of the model's performance across various classification thresholds.

A confusion matrix is utilized to provide an overview of a classification model's performance. This table shows information about the number of true positives, true negatives, false positives, and false negatives, which are generated by the model, highlighting its performance across different classes. This tool is particularly useful in evaluating the performance of models that deal with imbalanced datasets or multiple classes. It is a useful tool for evaluating model performance across different classes. Precision, recall, and F1 score are commonly calculated from the confusion matrix, providing further insights into the model's performance for different classes.

3. Result

3.1. Result of different structures of Bi-LSTM model

Based on the initial bidirectional LSTM model, other parameters are kept unchanged, multiple bidirectional LSTM layers are added, the grid structure is modified, the model is compiled and fitted, and different results about the prediction accuracy and loss rate of sarcasm analysis are obtained as Table 1. These results are plotted in the following table to explore the optimal LSTM grid structure. Confusion matrix and ROC results achieved from the optimal structures are presented in Figure 1 and Figure 2 respectively.

Table 1. Result comparison of various structures.

Bi-LSTM model	Accuracy	Loss
One layer	87.83%	29.92%
Two layers	87.62%	30.53%

DOI: 10.54254/2755-2721/21/20231118

Table 1. (continued).

Three layers	87.64%	30.40%
Six layers	86.95%	31.19%

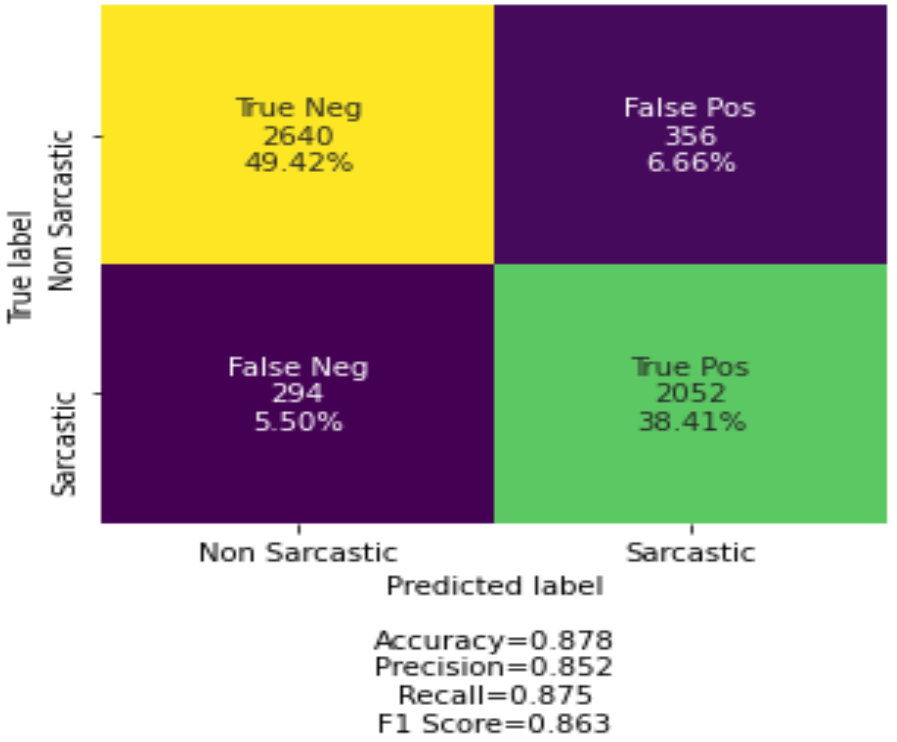

Figure 1. Confusion matrix results (Picture credit: Original).

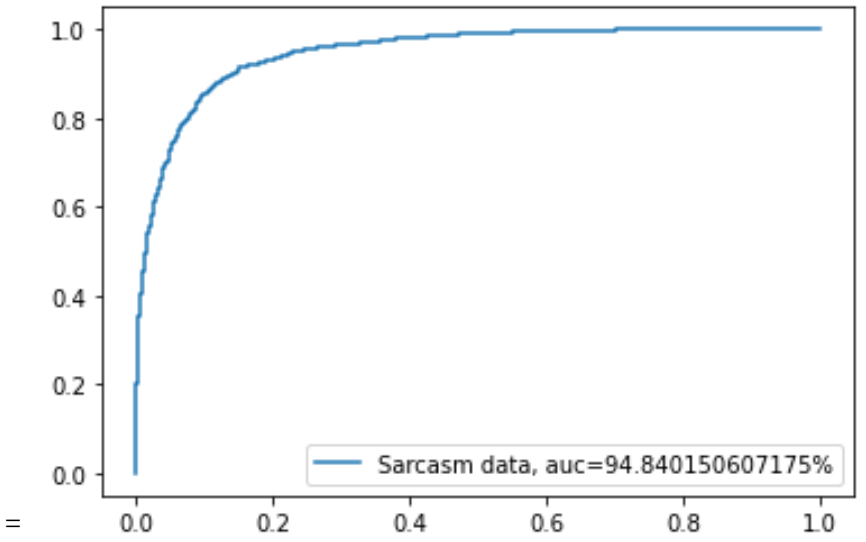

Figure 2. ROC curve result (Picture credit: Original).

It is worth noting that this table contains four different grid structures, namely one-layer, two-layer, three-layer, and six-layer bidirectional LSTM models. Among them, the purpose of first three layers deign is to gradually explore the effect of the quantity of layers in the model on the accuracy of the

DOI: 10.54254/2755-2721/21/20231118

forecast results. The reason for setting the six-layer model is that the accuracy of model prediction has not been improved from single layer to three layers. Therefore, a model with a larger amount of layers is set to observe whether there is some factor that makes it difficult to affect the accuracy of model prediction results by simply adding bidirectional LSTM layers.

3.2. Sarcasm prediction using BERT model

The accuracy, loss and AUC of the prediction results are compared, and figure 2 is obtained to explore the reasons for the difference in the prediction results between the Bi-LSTM and the BERT model, and the two models are compared and analyzed.

Table 2. BERT performances on sarcasm prediction.

	Accuracy	Loss	AUC
One layer Bi-LSTM model	87.83%	29.92%	0.9484
BERT model	91.84%	15.02%	0.9770

4. Discussion

In theory, modifying the LSTM model's grid structure and layer depth can give a rise to its complexity and ability to capture long-term dependencies within the input sequence, improving its performance in tasks such as language modeling, machine translation, and sentiment analysis. This is significant for natural language processing tasks where long sequences need to be modeled. At the same time, more LSTM layers can capture more information in the input sequence, which can enhance the generalization ability and the model accuracy. However, analyzing the results of previous studies, this work finds that simply adding multiple LSTM layers to make the grid structure of the model more complex can indeed increase the complexity and expressiveness of the model to a certain extent, but does not enhance the model prediction accuracy. Possible reasons could be as follows:

Firstly, data quality issues. There is noise, outliers, etc. in the dataset, the latent patterns are not enough to be learned by the model, and adding more LSTM layers will not bring significant improvement to the results.

Secondly, hyperparameter tuning issues. Not only the number of LSTM layers will affect the grid structure of the model, but also the selection of hyperparameters such as the size of LSTM layers, dropout rate, learning rate, number of epochs, and optimizer will affect the model performance. Not carefully tuning and optimizing these hyperparameters according to the actual situation may limit the performance of the model.

Thirdly, there is an overfitting problem. Adding more layers appends model complexity and causes the model to overfit the training data. Overfitting causes the model to learn noise and patterns in the training data, which affects its ability to generalize to new data.

Fourthly, diminishing returns. Adding more layers does not necessarily improve the model accuracy, as each additional layer brings progressively less benefit. Maennel and Schneider studied the expressive power of deep neural networks [9]. After increasing a certain number of layers, the model may suffer from the vanishing gradient problem, resulting in earlier layers failing to learn a useful representation of the input data.

When comparing the best prediction results of the Bi-LSTM model with that of the BERT model, it could be observed that the performance of the Bi-LSTM model is slightly lower than that of the BERT model when the same data set is used. There are several possible reasons:

Firstly, model Architecture issues. The architecture of bidirectional LSTM models is relatively not complex enough to handle the complexity of BERT models processing text. The BERT model uses the Transformer architecture and can handle longer sequences and more complex semantic information. In contrast, the Flatten layer in the bidirectional LSTM model may fail to capture the information in

DOI: 10.54254/2755-2721/21/20231118

longer text sequences, which may cause the model to lose some important semantic information during prediction.

Secondly, preprocessing issues. Although the bidirectional LSTM model using pre-trained GloveData word vectors as the embedding matrix, and this kind of approach does have enhanced the accuracy of the model in the previous experiments, these embedding vectors may not fully cover the language information in the pre-trained language model used by BERT model. BERT uses preprocessing techniques such as WordPiece embedding and masked language modeling to better process raw text data, which can help improve model performance and accuracy [10].

Thirdly, parameter setting issues. The parameter settings such as lstm_out and dropout, in the bidirectional LSTM model may need further adjustment. These parameters need to be adjusted appropriately according to the characteristics of the data set and the experimental results.

At the same time, this paper shows when using bidirectional LSTM model for sarcasm prediction, the accuracy of the model prediction is only nearly four percentage points lower than that of the BERT model, which may be caused by the complexity of the dataset features and the text task. The text length of many data in the experimental data set is short, so that the BERT model cannot play the advantage of the Transformer architecture when processing longer text sequences. At the same time, the text task of sarcasm analysis involved in this experiment is relatively simple, which only needs to predict whether the input text is sarcastic or not, which may not require a very complex model.

5. Conclusion

Through the exploration and assessment of experimental results, the following statements could be concluded. Firstly, by using the control variable method, results regarding different structures of the LSTM model: It is found that it is infeasible to change the model structure and enhance the model complexity by simply increasing the amount of LSTM model layers, so that the result prediction accuracy of the model is improved. There may be problems such as model overfitting, diminishing returns, hyperparameter adjustment, and data set quality. Secondly, by comparing the predicting results accuracy with the BERT model using the same dataset, it is found that the Transformer model has higher accuracy when dealing with long sequences and high complexity text information. At the same time, the preprocessing techniques used by the BERT model can better process the original text data and help to improve the model performance. Thirdly, observing little difference in the accuracy of sarcasm prediction results using bidirectional LSTM model and BERT model. It is found that when the length of the text in the dataset is short and the complexity of the text task is low, the accuracy of the BERT model and bidirectional LSTM model are close to each other.

In the future text analysis task of sarcasm prediction, when bidirectional LSTM model is leveraged for prediction, the hyperparameters could be adjusted of the model and optimize the dataset on the ground of changing the number of layers to improve the accuracy of the LSTM model in sarcasm prediction. Moreover, a more complex and more accurate model like BERT could be applied for text analysis prediction in related tasks. At the same time, for various types of sarcasm expressions and text data in different fields and languages that may appear in the future, more robust sarcasm prediction models could be explored, together with the method to incorporate text background and situation information into sarcasm prediction models to enhance the performance of predicting models, and further improve the performance and application value of sarcasm prediction.

References

[1] Luo, B., Lau, R. Y., Li, C., & Si, Y. W. (2022). A critical review of state-of-the-art chatbot designs and applications. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12(1), e1434.

[2] Sarsam, S. M., Al-Samarraie, H., Alzahrani, A. I., & Wright, B. (2020). Sarcasm detection using machine learning algorithms in Twitter: A systematic review. International Journal of Market Research, 62(5), 578-598.

DOI: 10.54254/2755-2721/21/20231118

[3] Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. Electronic Markets, 31(3), 685-695.
[4] Khanal, S. S., Prasad, P. W. C., Alsadoon, A., & Maag, A. (2020). A systematic review: machine learning based recommendation systems for e-learning. Education and Information Technologies, 25, 2635-2664.
[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[6] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
[7] Wang, S., Wang, X., Wang, S., & Wang, D. (2019). Bi-directional long short-term memory method based on attention mechanism and rolling update for short-term load forecasting. International Journal of Electrical Power & Energy Systems, 109, 470-479.
[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., et, al. (2017). Attention is all you need. Advances in neural information processing systems, 30, 5998-6008.
[9] Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., & Sohl-Dickstein, J. (2017). On the expressive power of deep neural networks. In international conference on machine learning, 2847-2854.
[10] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.