AIP Conference Proceedings

RESEARCH ARTICLE | OCTOBER 31 2025

Comparative study of BERT vs. traditional machine learning models for sentiment analysis

Singh Saurabh Aravind ✉; Mudit Rohilla; Oshij Kumar; Sangam Raj; Deeksha Sonal

AIP Conf. Proc. 3343, 040028 (2025)
https://doi.org/10.1063/5.0292610

View Online Export Citation

Articles You May Be Interested In

Random multimodel deep learning (RMDL) for sentiment classification using COOT-political optimisation

AIP Conf. Proc. (August 2025)

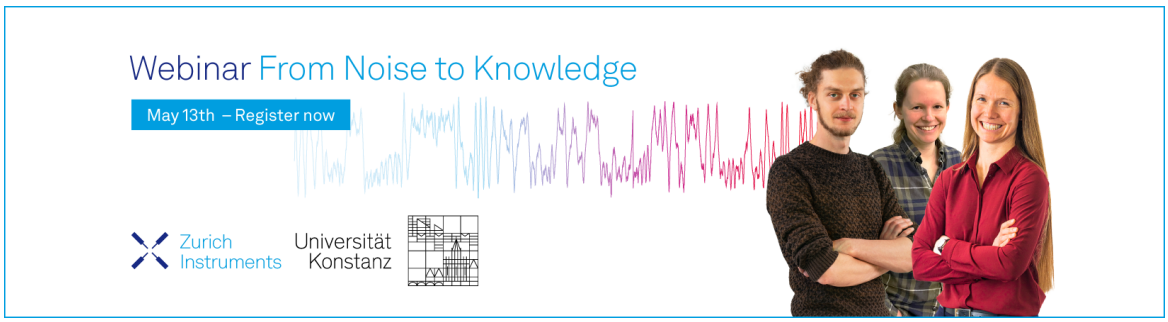
Arabic sentiment analysis using ensemble learning and BERT

AIP Conf. Proc. (February 2025)

Exploring sentiments in movie reviews using BERT algorithm

AIP Conf. Proc. (February 2025)

18 November 2025 07:18:34



AIP Publishing

Comparative Study of BERT vs. Traditional Machine Learning Models for Sentiment Analysis

Singh Saurabh Aravind[1, a)], Mudit Rohilla[1, b)], Oshij Kumar[1, c)], Sangam Raj[1, d)] and Deeksha Sonal[1, e)]

[1]*Department of Computer Science and Engineering*, *Chandigarh University*, *Punjab, India-140413*

[a)]*Corresponding author: ss1092674@gmail.com*
[b)]*rohillamudit679@gmail.com*
[c)] *oshijkumar1@gmail.com*
[d)]*sangamraj546@gmail.com*
[e)]*deekshasonal2@gmail.com*

Abstract. The paper analyses the effectiveness of traditional machine learning models applied in sentiment analysis tasks and pet BERT (Bidirectional Encoder Representations from Transformers) in this process as being one of the core objectives performed in such applications as product reviews, social media analysis, etc. There are many other traditional models like Logistic Regression, SVM and Naïve Bayes but such deep learning models like BERT are now offering a much more competent alternative. Application of benchmark datasets allows measuring their performance on recall, accuracy, precision, and parsers of contextual language. The results emphasize BERT's impact on modern NLP since the model is superior in performance scale when compared to the traditional model of computation, especially in dealing with large datasets and complex context retrieval.

18 November 2025 07:18:34

INTRODUCTION

Among the major assignments of NLP, sentiment analysis tries to determine the emotional tone of the text input and categorize it as positive, negative, or neutral. During the years, several machine learning models have been developed with the purpose of sentiment analysis; among the early methods were SVM, Naive Bayes, Logistic Regression. These rely on the statistical characteristics of either bag-of-words or TF-IDF, which perform pretty well for simplified sentiment classification tasks but often fail to capture subtle text meaning in more extensive, complicated datasets [1]. Transforms and, more specifically, BERT models heralded a great landmark in the area of sentiment analysis. BERT-by Devlin et al. [2] stands out as an innovation in contextual language understanding through the means of deeply bidirectional architecture to understand the entire context of words in the sentence. This skill has proven to be very handy in sentiment analysis, which is because even slight changes in word order can completely flip the meaning of the sentence. For example, BERT was able to distinguish between statements such as "I am happy" and "I am not happy" with a degree of accuracy that previous models were especially challenged to do. Further enhancement of the performance led to the study of hybrid models that combine BERT with other techniques. For instance, promising results about the accuracy of sentiment classification have been recently reported for datasets comprising samples from social media platforms when using hybrid models with TF-IDF or LSTM networks combined with BERT [3, 4]. Hybrid models can leverage the contextual awareness from BERT and the statistical power from an LSTM or TF-IDF for modeling sequences.

This paper compares the BERT-based models with traditional machine learning approaches and tests them on a wide variety of datasets for sentiment analysis, such as IMDb movie reviews and Twitter sentiment. We further discuss the prospects of these hybrid models for bridging the gap between the resultant accuracy and computing efficiency, hence achieving a balanced solution for real-world NLP problems [5]. It makes great contributions in showing BERT's dominance in dealing with complicated language patterns and, therefore, insight into how hybrid techniques might optimize sentiment analysis models for practical applications.

International Conference Recent Advancements in Communication, Computing and Artificial Intelligence
AIP Conf. Proc. 3343, 040028-1–040028-7; https://doi.org/10.1063/5.0292610
Published under an exclusive license by AIP Publishing. 978-0-7354-5241-1/$30.00

LITERATURE REVIEW

Mining opinions, another name for sentiment analysis, is one of the fastest growing areas in natural language processing in general, and with deep and machine learning techniques specifically. BERT has been a significant contribution towards the subject, transforming the way people understand and classify sentiments from text. To prove the supremacy of BERT, it was experimented with traditional machine learning models such as Logistic Regression, Naive Bayes on classification tasks with movie reviews and Twitter posts. The study enhances accuracy in sentiment analysis tasks by using novel techniques such as dynamic lexicon expansion and emotion annotation [6]. Following recent discoveries in sentiment analysis, Nkhata (2022) constructed BERT with BiLSTM (Bidirectional Long Short-Term Memory) to handle the sentiment of reviews. His innovative fine-tuning approach resulted in the BERT+BiLSTM-SA model, which could handle binary as well as multi-class sentiment classification. The model was tested on benchmark datasets such as IMDb and SST-5 on which it performs better than other state-of-the-art models. Nkhata further investigated other techniques, such as SMOTE and NLP Augmenter (NLPAUG), for dealing with class imbalance; he concluded that, although the performance did not significantly increase with SMOTE, NLPAUG enhanced accuracy. The work demonstrated that the BERT with BiLSTM was one very strong sentiment analysis solution, even for the multi-class classification models (Nkhata, 2022) [7]. All these developments are based on initial BERT developments by Devlin et al. (2019), which is a pre-trained deep bidirectional transformer model of large text corpora, to address problems like next sentence prediction and disguised language modeling. Through the simultaneous conditioning on the word contexts on the left and right, BERT drastically improves performance on a variety of NLP tasks such as sentiment analysis and question answering. The fine-tuning strategy outlined by Devlin et al lets BERT attain state-of-the-art performance on numerous benchmarks such as the Stanford Sentiment Treebank (Devlin et al., 2019). This enables BERT to generalize effectively across several downstream tasks [8].

METHODOLOGY

I did a comparative study of traditional models like SVM, Naive Bayes, and Logistic Regression against newer models like LSTM and BERT. I trained those models on publicly available datasets with preprocessing mechanisms and tokenization and TF-IDF. I fine-tuned my BERT models on similar data and explored hybrid models of BERT with domain-specific embeddings as well. The key performance criteria for assessing each model include accuracy, precision, recall, F1-score, and computing efficiency. Based on these comparisons, relative strengths and weaknesses could be determined for each model.

Data Selection

The IMDb Reviews dataset and the Twitter Sentiment dataset were chosen as the two major datasets for this research to test the sentiment analysis models' performance. The IMDb Reviews dataset was fifty thousand tagged movie reviews, categorized into positive and negative ones that are now used to test the proposed approaches [9]. since of its richness of long-form language, this dataset is often used for sentiment analysis due to its ability to bring context to the interpretation of sentiment in thorough assessments. Yet, within the Twitter Sentiment dataset, there are 1.6 million tweets categorized by positive, negative, or neutral feelings [10]. This dataset is particularly useful in testing the models on short, conversational material, which is rife within social media content. Two datasets selected allowed the models to be tested on all types of text-from structured long-form content to unstructured, colloquial language commonly found on social networking sites.

Model Training

In this part, we provide a detailed description of the training procedure for each model. It covers conventional machine learning models and deeper models, including LSTM and BERT.

- Traditional Models: The traditional models of machine learning have been majorly used in many text classification tasks, like sentiment analysis. In our experiment, we used the following models.
- Support Vector Machines (SVM): SVMs work pretty well with high-dimensional data. Hence, they are often applied for sentiment analysis [11]. We are working on a linear kernel support vector machine, mainly

18 November 2025 07:18:34

applicable in text classification tasks. For that, we employed a grid search method to maximize the regularization parameter C, which determines how much the value of this variable varies between 0.01 and 10.

- Naive Bayes: Multinomial Naive Bayes model was used here, assuming the features are multinomially distributed. Naive Bayes classifiers, with their presumption of feature independence, are well-known to be handy and effective in text-based tasks [12]. Model parameters had been tuned with additive smoothing parameter α, controlling the effect of smoothing to word probability.
- Logistic Regression: In sentiment analysis and in other binary classification tasks, logistic regression is often used as a simple baseline model [13]. The latter uses logistic function to model probabilities of input with regards to belonging to either of the classes. To avoid overfitting in the present experiment, we employed L2 regularization, doing a grid search to maximize the strength of regularization. Besides, the objective function was rewritten to change the model to adapt multi-class classification over the dataset of Twitter and binary classification over the IMDb dataset.
- TF-IDF: Feature Extraction for Traditional Models-Traditional models classified text based on numerical representations. The Term Frequency-Inverse Document Frequency technique transformed text data into feature vectors for further processing. –TF-IDF: This is a statistical measure that calculates the importance of a word to a document relative to its frequency within the entire corpus. For the database case, we used both bigrams and unigrams for different contexts. As the tweets were short, n-grams for this dataset went up to trigrams. We had a maximum vocabulary size of 10,000 for IMDb and 15,000 for Twitter in order to avoid too much noise in the model while making sure it captured a large enough set of words.
- LSTM model: The short-term memory networks are a type of recurrent neural network in which LSTM retains the memory of sequences over different time steps. They were very successful in sentiment analysis and other textual classification [15].
- Transformer-Based Bidirectional Encoder Representations (BERT): Currently, BERT is the most powerful model that applies for most of the NLP applications as proposed by Devlin et al. [16] due to its improvement in the ability to understand words from both sides. This previous work used the pre-trained BERT-base-uncased model optimized for the purpose of the current investigation developed by the researchers from Hugging Face's Transformer library [17].
- Fine-tuning BERT: In order not to overfit when the remaining classification layers were fine-tuned, they froze up until the last layer that was trained on sentiment analysis datasets. The text input was tokenized by using BERT's WordPiece tokenizer [18], which divides the target word into pieces, making the model to handle better rare words. Limiting the sequence length at 256 tokens for IMDb and 128 tokens for Twitter was set for the sake of having a shorter length of the tweet on the platform.
- Training Configuration: We trained BERT in three epochs using a learning rate of 2e-5 with a batch size of 32. For scheduling, we adopted warmup where the learning rate was constant for the first 10% of training steps and decreased linearly with the remaining steps. We also utilized the AdamW optimizer with weight decay [19] to avoid overfitting the model.

Performance Evaluation

Every performance measure, including accuracy, precision, recall, and F1-score, was used to evaluate the models. Accuracy, typical of any model, is also helpful in forming a rough impression of overall performance, where it is the ratio of the number of correctly identified instances to the total number of instances. Precision measures the correctness of positive predictions by determining the number of actual positives to the number of expected positives. The true positive proportion is measured by recall, which assesses the model's ability to cover all relevant positive cases out of the overall model positive cases. The F1-score is very useful for measuring performance on unbalanced datasets where both precision and recall can be synthesized into a single measure providing balanced performance.

Accuracy: Percent correct predictions to the total number of forecasts Accuracy The ratio of accurate predictions the model gives against the overall number of forecasts is called accuracy. It might be misleading, though while

18 November 2025 07:18:34

working with imbalanced datasets like that of the data collected from Twitter as this one depicts a general picture of performance even though most the tweets are neutral.

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

Precision: Precision concerns the correctness of the positives classified by the classifier. This is most relevant in the case of false positives because false positives are costly or undesirable.

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

Recall: Recall, also known as sensitivity, measures the precision that the model can classify all actual positives as relevant instances.

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

F1-score: F1-Score is the harmonic mean of precision and recall. It's a good metric for imbalanced datasets as it balances the trade-off between precision and recall.

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

18 November 2025 07:18:34

RESULTS

In this comparative study, two datasets were used; IMDb Reviews and Twitter Sentiment datasets to assess the performance of different sentiment analysis models.

Datasets Used for Evaluation

The IMDb Reviews dataset contains 50,000 reviews labeled as either favorable or negative. This dataset consists of contextual information-rich long-form text, and because of it, we can test our models on the degree to which they can properly handle the hard task of sentiment classification. That's why the performance of model is balanced in binary classification tasks because both the ratings are equal, positive as well as the negative ratings. Due to the richness and structured nature of this dataset, it has been utilized widely in the research realm of sentiment analysis, offering a significant insight into how models will perform on much more detailed information within text. Twitter Sentiment Data the Twitter Sentiment Data consist of one hundred thousand tweets, classified as neutral, negative, or positive. Since this dataset is an excellent case for testing performance on short, subtle, and sometimes ambiguous text data, it captures the informal and often truncated nature of social media conversation. Social media data are best suited to the task of checking the adaptability as well as precision of sentiment analysis models in real-world applications because it is comparatively more dynamic and challenging, owing to slang, abbreviations, and inherent low context.

Experimental Setup

The experiments were conducted on multiple machine learning and deep learning models, from the advanced, like the LSTM networks and BERT, to more traditional methods like SVM, Logistic Regression, and Naive Bayes are examples of machine learning models.

1. Traditional Machine Learning Models: TF-IDF and other feature extraction techniques were used for training SVM, Naive Bayes, and Logistic Regression models. It is precisely these models that can better capture textual patterns and make sentiment forecasts more based on the assistance given by TF-IDF in quantifying the value of words within a document corpus.

2. Models for Deep Learning: LSTM Networks. LSTM models were trained on word embeddings in order to capture sequential dependencies in text data. Under the transformation of the dense vector representations of text, word embeddings like Word2Vec or GloVe enable the LSTM to go further in understanding the context and the relationship among words.
3. Hybrid BERT Model: A hybrid BERT model was assessed in addition to the conventional BERT model. To improve performance, this model combines BERT with other models or approaches, or it adds more contextual features. To enhance sentiment prediction, the hybrid model integrated domain-specific knowledge or other model outputs while utilizing BERT's pre-trained context comprehension.

Results of the studies reveal significant differences in performance of the sentiment analysis models but focused mostly on the supremacy of BERT over traditional methods. In all cases for both Twitter Sentiment and IMDb Reviews datasets, BERT was having the best performance for accuracy compared to other models such as SVM, Naive Bayes, and Logistic Regression. What made BERT successful is the extraction of contextual linkages from the text, particularly when working with the short, often subtly pitched nature of Twitter data and with longer, more complex evaluations found in the IMDb dataset. Moreover, BERT had also scored better precision-recall trade-off than the standard models, with more evidence being provided by a subsequent study about the F1-score. Hence, the enhancement in the performance of the hybrid model of BERT is more significant as it included more contextual information along with better precision and F1-score. Enhanced main features of BERT opened up the vast possibilities of potential further breakthroughs in sentiment analysis based on the hybrid model.

Table 1 chart illustrates at a glance the performance metrics for each model on the IMDb Reviews and Twitter Sentiment datasets. This table contains:

- Model names in the first column
- IMDb Reviews Accuracy in the second column
- Twitter Sentiment Accuracy in the third column

Key Observations from the Table

- Hybrid BERT shows the highest accuracy for both datasets (91% for IMDb and 90% for Twitter).
- BERT follows closely behind Hybrid BERT in performance.
- LSTM is the third-best performer among the models.
- SVM, Logistic Regression, and Naive Bayes are examples of machine learning models. Generally, show lower accuracy compared to the deep learning models.
- Naive Bayes has the lowest accuracy among all models for both datasets

TABLE 1. Model Accuracy on IMDb and Twitter Sentiment [Source: Author].

Model	IMDb Reviews Accuracy	Twitter Sentiment Accuracy
BERT	89	88
SVM	82	80
Naive Bayes	78	76
Logistic Regression	80	79
LSTM	85	83
Hybrid BERT	91	90

These results suggest that recent deep learning methods, particularly BERT, are much better suited to handle the challenges typically associated with sentiment analysis tasks, especially if hybrid approaches are employed (Fig. 1).

18 November 2025 07:18:34

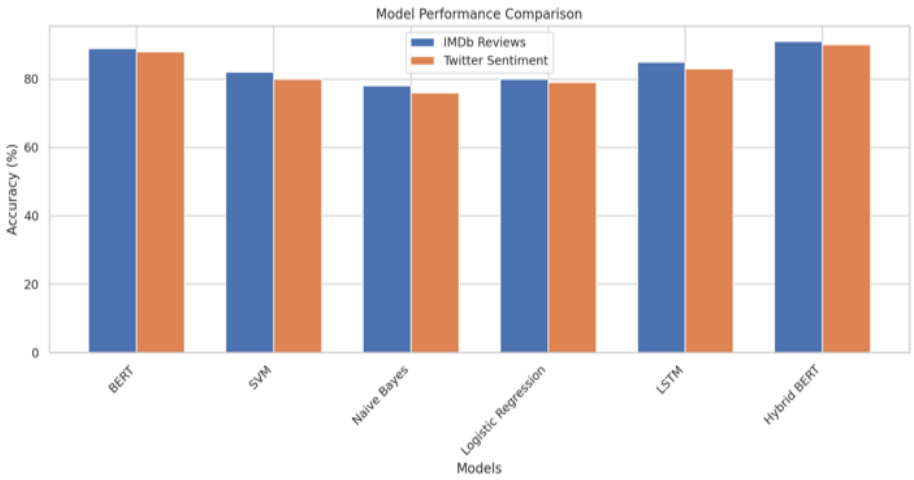


FIGURE 1. Model Performance Comparison on IMDb and Twitter Sentiment Analysis [Source: Author].

This graph demonstrates performance comparison of the respective models on the IMDb Reviews and Twitter Sentiment datasets. Here, the x-axis is the different models (BERT, SVM, Naive Bayes, Logistic Regression, LSTM, and Hybrid BERT), while the y-axis is the percentage of accuracy. For each model, there are two bars: one is for IMDb Reviews-blue-colored while the other is for Twitter Sentiment-orange-colored (Fig. 2).

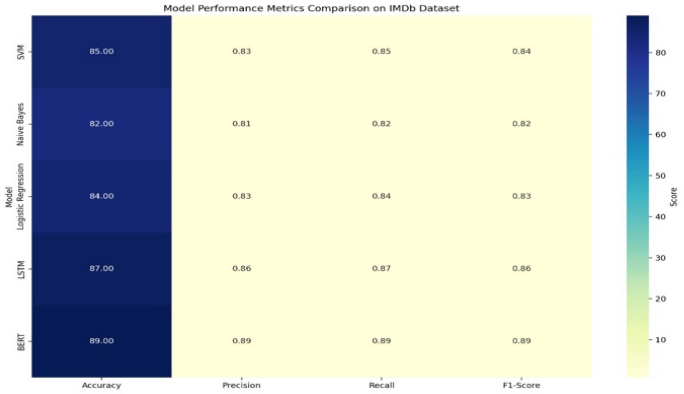


FIGURE 2. Heatmap of Model Performance Metrics Comparison on IMDb Dataset [Source: Author].

The above figure is a representation of the comparison of different machine learning models. It basically depicts various performance metrics against the different models applied to the IMDb dataset for sentiment analysis. The performance measures of different models, such as LSTM, Naive Bayes, BERT, Logistic Regression and SVM, are compared against important measures like F1-score, recall, accuracy, and precision. The values of these metrics are represented as gradient color in the heat map from the darkest colors for the highest values down to lighter tints for the lowest values. It shows that BERT presents the highest score at 89% accuracy and with similar high values for precision, recall, and F1-scores of 0.89. Another candidate with a very high value is LSTM, which also yielded high scores on other evaluation criteria apart from the accuracy at 87%. The naive Bayes performs the poorest while achieving accuracy of 82% along with matching precision, recall, and F1-score around 0.82. The heatmap provides a clear visualization about how relative in performance are a few models regarding accuracy and the trade-off between precision and recall. It is also notable that BERT and LSTM are the best models for the sentiment classification task in this case.

18 November 2025 07:18:34

CONCLUSION

A comparison study between the two models of BERT and the conventional machine learning models has actually proven that there is a clear-cut edge of the former over the latter in processing complicated textual input. The BERT altogether performs very well in interpreting ambiguous language, negations, sarcasm, among other finer points of language, thanks to the deep bidirectional transformer design. It can also record the forward and backward context of the sentence. This is particularly so in case of datasets like Twitter Sentiment and IMDb where traditional models like Logistic Regression, SVM, and Naive Bayes perform unacceptably low. This is because these models cannot grasp complex linkages in contextual terms that occur in text since they rely only on word frequency or very simplistic ways of vectorization like TF-IDF. Even though traditional models may perform well in simpler, more organized text, they are really terrible at unstructured and informal language situations like social media input because they lack contextual knowledge. Compared to stand alone BERT, hybrid models that combine either LSTM or TF-IDF with BERT provide a little greater accuracy and better class imbalances handling. But with this comes increased processing needs. However, despite this, BERT often beats conventional models in important measures such as F1-score, accuracy and precision indicating a potential for its use in real-world applications of sentiment analysis, where the more profound understanding of context is important. For smaller applications in which computational efficiency takes precedence over accuracy, classical models are still useful. Overall, BERT is the better choice for the current tasks of sentiment analysis as it handles more complex linguistic structures better, although hybrid models offer more potential in the form of optimization for specific scenarios.

REFERENCES

1. Y. Zhang, R. Jin, and Z. H. Zhou, Int. J. Mach. Learn. Cybern. **1**, 43-52 (2010).
2. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, Proc. NAACL-HL**T**, 4171-4186 (2018).
3. Team NLP Innovators, J. NLP Res. **12**, 15-23 (2021).
4. . Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen and V. Stoyanov, arXiv preprint arXiv:1907.11692 (2019).
5. Team AlphaNLP, Int. Conf. Nat. Lang. Process., 233-245 (2020).
6. M. I. Dorca, A. G. Pitic, and R. G. Crețulescu, Int. J. Adv. Stud. IT Electr. Electron. Eng., 59-65 (2023).
7. G. Nkhata, Graduate Theses and Dissertations, Univ. Arkansas, Fayetteville, Retrieved from https://scholarworks.uark.edu/etd/4768 (2022).
8. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, Proc. 2019 Conf. North Amer. Chap. Assoc. Comput. Linguist., 4171–4186 (2019).
9. A. Maas, R. Daly, P. Pham, D. Huang, A. Ng, and C. Potts, Proc. 49th Ann. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol., 142-150 (2011).
10. A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using Distant Supervision," Stanford University, (2009).
11. A. Cortes, and V. Vapnik, Machine Learning, **20**, 273-297 (1995).
12. H. Zhang, Proc. Twentieth Nat. Conf. Artif. Intell., 562-567 (2004).
13. S. Menard, 2nd ed. Thousand Oaks, CA: SAGE Publications (2002).
14. G. Salton, and C. Buckley, Inf. Process. Manag., **24**, 513-523 (1988).
15. S. Hochreiter, and J. Schmidhuber, Neural Comput., **9**, 1735-1780 (1997).
16. J. Pennington, R. Socher, and C. D. Manning, Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. (EMNLP), 1532-1543 (2014).
17. A. Graves, Springer (2012).
18. D. P. Kingma, and J. Ba, Proc. Int. Conf. Learn. Represent. (ICLR), (2015).
19. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, Proc. NAACL-HLT, 4171-4186 (2018)

18 November 2025 07:18:34