Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Contextual Sentiment Boosting through Lexicon Masking and Transformer Fine-Tuning

Dr. Deepa D[1], Dr. Mehfooza M[2], Dr. S Prabhu[3], Dr. Padmavathy T[4]

[1,4]Assistant Professor (Sr.G-1), School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamilnadu 632014, India
[2]Associate Professor, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamilnadu 632014, India
[3]Associate Professor, Department of Computer Science and Engineering (Cyber Security), Nandha Engineering College, Erode, Tamilnadu 638052, India

ARTICLE INFO

Received: 31 Dec 2024

Revised: 20 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

Introduction: In today's advert environment sentiment analysis plays critical for gathering insights from consumer perspectives. Typical machine learning models rely on predefined features which might struggle with feature weighting, neglect important comments and misunderstand word meanings particular to a certain domain especially when applied to small datasets.

Objectives: This study aims to overwhelm the boundaries of traditional sentiment analysis by enhancing feature representation and leveraging deep learning. Also it focusses on improving the extraction of sentiment-related features, domain adaptation and optimizing model parameters for better performance.

Methods: The proposed approach includes fine-tuning a BERT with a focus on sentiment-appropriate terms by boosting the fraction of masked words during training, which allows more effective bidirectional contextual learning. Particle Swarm Optimization (PSO) is used for hyperparameter tuning to optimize model performance. Additionally, character-level and subword embeddings are used to handle unknown terms. Transfer learning is applied to enrich classification by integrating domain-adapted features.

Results: The PoS Masking-PSO BERT model achieved 96.6% accuracy on hotel and 95.2% on movie reviews, with F1-scores of 95.6% on both. Contrated to baseline BERT which have 89%, 93.87% performance improved significantly. Processing times dropped to 9 and 13 minutes. Optimal results used 12 encoder layers and 750 hidden units. PoS masking enriched feature context while PSO enhanced model stability and accuracy.

Conclusions Combining PoS masking and PSO into BERT strengthen sentiment analysis, competently addressing domain variation and feature weighting. The model surpasses existing BERT approaches in accuracy and F1-score and reduced processing time prompting it suitable for real-time applications. This modified architecture describes deep contextual features posing a robust solution for cross-domain sentiment analysis.

Keywords: Sentiment Analysis, BERT, Part-of-Speech Masking, Particle Swarm Optimization, Hyperparameter Tuning, Text Classification, Transfer Learning, Lexicon Masking, Bidirectional Transformers, Feature Representation.

INTRODUCTION

Sentiment analysis involves labeling emotions conveyed through text. Beyond a simple identification of positive or negative words Artificial Intelligence (AI) comprehends the overall sentiment of a text [1]. Organisations use sentiment analysis to estimate industry trends, improve customer service, control reputation, research rivals and improve marketing tactics [2]. The rise in technological progress entered with the plenty of online content has underscored the importance of comprehending human sentiments and opinions. This interaction has prompted the broad integration of sentiment analysis across a range of industries and uses [3]. Understanding the nuances of word meanings which can change based on the context is essential to accurate sentiment analysis [4]. Models may have trouble in correctly rendering sarcasm, irony, and other figurative language. In such cases a more thorough comprehension of the background and cultural indicators can be necessary. Due to variations in language, the model efficiency is reduced when a model trained in one domain may struggle in another, reducing its effectiveness [5].

User-generated content adds complexity entitled to slang, mixed languages, spelling errors, and professional jargon, making it difficult for standard NLP techniques to accurately identify key attributes like sentiment, intent or even the dominant language. These limitations hinder the models' ability to generalize across different text types leading to reduced accuracy in real-world applications [6]. To overcome the sentiment analysis obstacles, it is required to adapt rigid data collection, feature engineering, model building, fine-tuning and continued validation.

Also the Natural Language Proessing (NLP) models are to e build to efficiently handle distinct linguistic differences and confirm greater generalisation across various text types. Pre-trained language models like BERT, GPT, XLNet, and T5 have shifting NLP from traditional supervised learning to pre-training followed by fine-tuning [7]. This development has significantly improved models' context and leading to more precise and human-like responses in a variety of NLP applications [8].

Traditional sentiment analysis techniques subject to comprehend context-dependent meaning and textual semantic variances. The traditional methods like Word2Vec and GloVe used static word embeddings, giving each word a single representation independent of context [9]. These models also limit their comprehension of complex meanings by seeing words as distinct entities and failing to recognise the relationship between words in a sentence [10]. However, by leveraging context learning and dynamic embeddings contemporary pre-trained transformer models such as BERT, GPT, XLNet, and RoBERTa have proven to perform better and improving the accuracy of sentiment categorisation [11].

Understanding sentiment nuances requires context learning, especially when sarcasm, negation, or ambiguous language are involved. Because pre-trained models incorporate context-aware word representations, they greatly outperform conventional techniques [12]. GPT uses autoregressive modeling, BERT uses masked modeling, and XLNet uses permutation modelling [13]. XLNet is the most computationally expensive, BERT is lighter than GPT, making it the preferred choice for the experiment has state of art results in most of the NLP.

PROBLEM DEFINITION

Although many advancements in sentiment analysis using models like BERT were introduced, still several challenges remain unresolved. Traditional sentiment models struggle to capture context, sarcasm, and nuanced expressions due to their static embeddings. The contextual understanding feature of BERT using it's 15% of random token masking strategy can mask sentiment-bearing words such as adjectives and nouns. Moreover, manually choosing hyperparameter for tuning model is complex and can lead to suboptimal model performance. The recent BERT-based models also surface in domain adaptation and handling diverse, user-generated content.

To address these constraints, the following approaches are proposed to enhance sentiment analysis task:

- Improve BERT's contextual understanding by masking sentiment-bearing parts of speech such as nouns, verbs, adjectives and adverbs instead of randomly selecting tokens.
- build up sentiment representation by avoiding simultaneous masking of multiple sentiment-bearing words and incorporate n-gram context to preserve neighboring sentiment cues during word embedding
- Fine-tune the BERT model using multiple swarm intelligence algorithms to optimize encoder depth, attention heads, and hidden layer sizes.
- Expand sentiment analysis from binary classification to fine-grained categories like weakly positive or strongly negative.
- Validate the model on real-world datasets such as movie and hotel reviews using accuracy and F1-score as key performance metrics.

LITERATURE SURVEY

Using the BERT model [15] investigated the role of fine-tuning in eight distinct text categorization datasets. Three sorts of fine-tuning methods were employed in this. The first method is to choose the best classification layer and optimizer. The second technique is to use the target data to train the BERT model. The third technique is to conduct several tasks, such as optimizer selection and target data training. When compared to the other strategies, multi-task tuning gave the best results of the three. The fine-tuning method was carried out by [16] by pre-training the BERT model with triplet as the input sequence. This method was put to the test on three separate datasets: WN11, FB13, and UMLS. BERT's knowledge-based tweaking resulted in improved performance. [17] used the Grid-search technique to analyse the fine-tuned BERT in their study. The researchers discovered that adequ17ate bias correction, learning rate, and epoch selection enhanced BERT performance but did not decrease data forgetting.

The BERT model was pre-trained using the user's 1.5 million unlabeled datasets. The data normalisation procedure in the BERT models was improved because of the use of several medical-based corpus datasets. A work on fine-tuning for medical dataset categorization was released by [18]. [19] identified the terms in the phrase based on their relationship. The BERT model was trained to select the best words in a phrase using these relationships. The relational idea was further used to three datasets, including analogy test google and large analogy. To address this issue, [20] suggested the RoBERTa model, which included a dynamic masking task and did away with the BERT's next sentence prediction job. To increase contextualization, the pre-training collection of sentences is repeated ten times, and all the sentences are utilized in adaptive masking assignments. On typical datasets, they displayed the state-of-the-art outcomes.

[21] trained the BERT model using semantic relations. Microsoft and Quora datasets, as well as the semantic dataset and the Quora question pair dataset, were employed in the testing. The recommended semantic link provided the greatest outcomes in the SQUAD 2.0 dataset since it had an F1 score. Syntactic knowledge in BERT was largely encoded in the lower layers, whereas semantic knowledge is dispersed over the 12 layers. Only the final layeris tailored for the masked word prediction job. Therefore, the POS and word match accuracy of the intermediary layers are reduced. The layer-wise localization of linguistic knowledge, encoded at distinct layers for words of different syntactic categories, was discovered by [22]. The fine-tuning method was carried out by [23] by pre-training the BERT model with triplet as the training data.

According to [24], the data was pre-trained using three training approaches utilising the BERT model. Hybrid algorithm outperformed all other algorithms. In order to classify tweets according to their emotional content, [25] utilised the BERT model. By eliminating emotion symbols and frequent terms in the phrases, it pre-processed the data. Its F1 score increased by 7% compared to the other models. To obtain more detailed contextual information about the tokens, researchers experimented with the Masking task in several ways. Sentiment Knowledge Enhanced Pre-training (SKEP) was proposed by [26] to generate the embedding for the sentiment word by masking automatically derived sentiment words. A network named as BertMasker introduced by [27] which explicitly masks domain-related words to enhance domain-invariant sentiment features and improved multi-domain sentiment classification by 0.94% and cross-domain settings by 1.8%. [28] aims Factorized Masked Language Modeling (FMLM) to separate semantic aspects like sentiment, syntax, and topic in transformer models. By selectively masking these factors during training, the model learns more understandable and integrated depictions. Such disentangled learning improves downstream tasks requiring fine-grained semantic understanding and control. Latest improvements [29] in MLM propose effective methods. Improved pre-training is done by dynamically changing masking ratios rather than fixed BERT 15% and curriculum-based token selectio.

DiffusionBERT porpose by [30] which aids discrete diffusion processes into MLM. And the model learns to foresee a noisy diffusion process. It improves text generation quality through enhanced semantic coherence by recovering text from an absorbing state, which is the state the process reaches the maximum corruption. This method beats existing models in perplexity and BLEU scores. New advances in MLM examine improved masking techniques and model elements to boost domain adaptability and language understanding. Iterative Mask Filling is a text augmentation method proposed by [31] that works well for all NLP applications by persistently masking and refilling words with BERT's predictions. Selective Masking based introduced by [32] is based on Genre and Topical Characteristics. This model adapts ranking and masking key words according to content significance. BERT Pretraining Decoder]33] suggests a decoder-enhanced pretraining technique that progresses model performance without changing BERT's encoder and promising interoperability with existing frameworks. A BERT-based model PoemBERT introduced by [34] utilizing a classical Chinese poetry corpus. The sentiment and pinyin embeddings were incorporated into the model by recognizing the unique emotional depth and linguistic precision of poetry. Thus, the model is enhancing its sensitivity to emotional information and address challenges inquired by the phenomenon of multiple pronunciations for the same Chinese character. Also, the mode model employs Character Importance-based masking and dynamic masking strategies, significantly augmenting its capability to extract imagery-related features and handle poetry-specific information.

The BERT model is fine-tuned with the hyper-parameters: 12 Transformer blocks, a 768 hidden layer, 12 self-attention heads, 0.1 dropout probability, 4 epochs, a 2e-5 initial learning rate, and 24 batch sizes. The BERT-pair-QA-B model has the highest accuracy, 95.6%, of all the models they tested. Good hyper-parameter selection and pre-training in the future may enhance the effectiveness of NLP approaches [35]. [36] done an investigation of the BERT hyper-parameters fine-tuning procedure. They first reviewed the causes of data forgetting before analyzing how proper bias correction, learning rate, and epoch selection could enhance BERT performance and lessen data forgetting and instabilities discovered during routine fine-tuning. Here, tweaking was done through a grid-search methodologyThe optimal hyper-parameters were selected for each of the . [37] [38][39] proposed models using the grid search technique.

CURRENT METHODOLOGY

[40]] evaluated the text categorization method on a variety of diverse datasets such as a movie and a hotel review, as well as tweets and Portuguese news. They utilised a machine-learning method based on word frequency and inverse document frequency to identify the most relevant information. The BERT model is the alternative option. There are two steps to the pre-training process. For example, you could conceal the tokens and then use them to make an educated guess at the label based on them. Based on the varied datasets, it was then necessary to fine-tune the model of the BERT. Different machine learning methods such as H2OautoML, Ridge Classifier and Logistic Classifier were compared to BERT's prediction performance. Across all datasets, the BERT model was shown to be more accurate and had a better Kaggle score than the others.

The drawbacks faced in this approach is as follows:

- It performed only 15% of masking for training the BERT that misses different words.
- It also suggested that their approach can be improved with the help of hyperparameter tuning to improve its performance.

By using swarm algorithms such as particle swarm optimization and genetic algorithms, the prior work was able to solve the hyperparameter tuning problem. In order to test it, we used the hotel and movie reviews databases. According to the results of the performance evaluation, particle swarm optimization is the most effective method for tweaking the model's hyperparameter. It does not, however, explain the function of masking in fine-tuning the BERT's performance. In this paper, Parts of speech-based masking are used here to fine-tune BERT and particle swarm optimization is used for hyperparameter tweaking.

PROPOSED MRYHODOLOGY

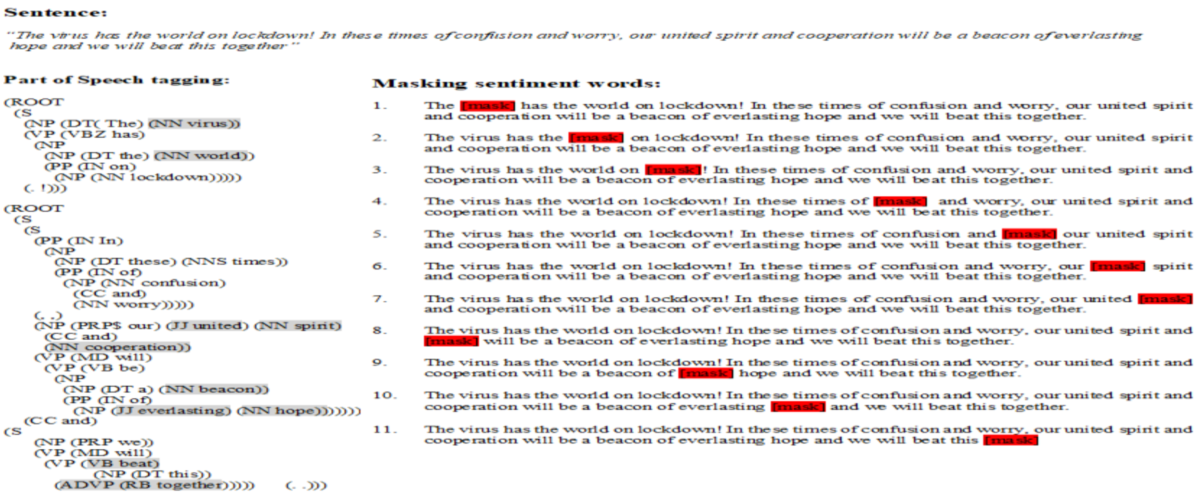
The BERT loss function in the classic BERT model only considers the prediction of masked values and ignores the prediction of non-masked words. As a result, the model converges more slowly than directed models, although this is counterbalanced by its greater context awareness. Based on the context supplied by the other, non-masked words in the sequence, the model then attempts to estimate the original value of the masked words. As a result, categorization performance suffers. To address this, a PoS-BERT model is presented, coupled with particle swarm-based hyperparameter tweaking, to improve the model's performance over the original BERT model.

4.1 Parts-of Speech - BERT model

In order to enhance sentence prediction from the BERT model, the Parts of speech-BERT tags nouns in a sentence as masked words. However, for specialist linguistics tasks, datasets are often less than 100MB, while BERT is trained on enormous unlabelled text datasets such as BooksCorpus and English Wikipedia, each of which contains 13GB of plain text Our semi-supervised learning method relies on a pre-trained linguistics model to categorise data from BooksCorpus and English Wikipedia. To learn POS tags, the pre-trained model learns them all at the same time.

4.1.1 Masking strategy

A Masked Language Model and Next Sentence Prediction (NSP) based on WordPiece embeddings are utilised in BERT. Using the training data generator, BERT selects 15 percent of the token locations at random for mask replacement and predicts masked tokens for Masked LM training goal. In this, the whole word masking scheme is used to mask the words. But it masks the important nouns in the sentence using Parts of speech tagging scheme instead of 15% masking scheme. the sample masking strategy for a sentence using the PoS-BERT is shown in figure 1.



Sentence:

"The virus has the world on lockdown! In these times of confusion and worry, our united spirit and cooperation will be a beacon of everlasting hope and we will beat this together"

Part of Speech tagging:

```
(ROOT
  (S
    (NP (DT( The) (NN virus))
    (VP (VBZ has)
      (NP
        (NP (DT the) (NN world))
        (PP (IN on)
          (NP (NN lockdown)))))
    (. !)))

(ROOT
  (S
    (S
      (PP (IN In)
        (NP
          (NP (DT these) (NNS times))
          (PP (IN of)
            (NP (NN confusion)
              (CC and)
              (NN worry)))))
      (, ,)
      (NP (PRP$ our) (JJ united) (NN spirit)
        (CC and)
        (NN cooperation))
      (VP (MD will)
        (VP (VB be)
          (NP
            (NP (DT a) (NN beacon))
            (PP (IN of)
              (NP (JJ everlasting) (NN hope)))))))
    (CC and)
(S
      (NP (PRP we))
      (VP (MD will)
        (VP (VB beat)
              (NP (DT this))
          (ADVP (RB together)))))     (. .)))
```

Masking sentiment words:

1. The [mask] has the world on lockdown! In these times of confusion and worry, our united spirit and cooperation will be a beacon of everlasting hope and we will beat this together.
2. The virus has the [mask] on lockdown! In these times of confusion and worry, our united spirit and cooperation will be a beacon of everlasting hope and we will beat this together.
3. The virus has the world on [mask]! In these times of confusion and worry, our united spirit and cooperation will be a beacon of everlasting hope and we will beat this together.
4. The virus has the world on lockdown! In these times of [mask] and worry, our united spirit and cooperation will be a beacon of everlasting hope and we will beat this together.
5. The virus has the world on lockdown! In these times of confusion and [mask] our united spirit and cooperation will be a beacon of everlasting hope and we will beat this together.
6. The virus has the world on lockdown! In these times of confusion and worry, our [mask] spirit and cooperation will be a beacon of everlasting hope and we will beat this together.
7. The virus has the world on lockdown! In these times of confusion and worry, our united [mask] and cooperation will be a beacon of everlasting hope and we will beat this together.
8. The virus has the world on lockdown! In these times of confusion and worry, our united spirit and [mask] will be a beacon of everlasting hope and we will beat this together.
9. The virus has the world on lockdown! In these times of confusion and worry, our united spirit and cooperation will be a beacon of [mask] hope and we will beat this together.
10. The virus has the world on lockdown! In these times of confusion and worry, our united spirit and cooperation will be a beacon of everlasting [mask] and we will beat this together.
11. The virus has the world on lockdown! In these times of confusion and worry, our united spirit and cooperation will be a beacon of everlasting hope and we will beat this [mask]

Figure 1. Proposed masking strategy

Parts of speech tagging is used in Figure 1 to mask all nouns in the phrase. In the BERT model, the number of masked words has grown because of this, and it enhances the next sentence prediction.

4.1.2 PoS-BERT Architecture

In this the architecture consists of four modules called a token generator, transformer, PoS tagging module and decoder. Here, multi-task learning is utilized to share the learning between the modules and PoS tagging modules. The training is based on both the loss minimization of language model in transformers and PoS tagging modules.

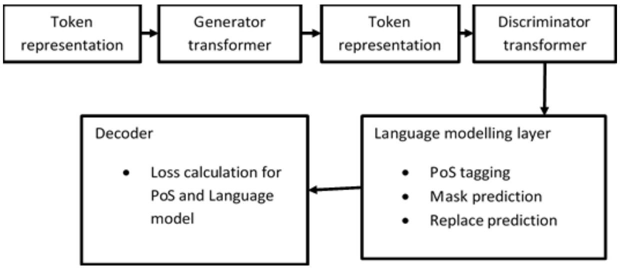


Figure 2. Proposed BERT architecture

a. Token formation

It's always the [CLS] token that comes first. A unique token [SEP] is used to separate input X from the sentence pair X1;X2. Each token's embedding vector in the Transformer encoder is a sum of its word, segment and positional embedding vectors. Since the language model is only applied to the final token from the transformer, both the token and annotation lengths are the same.

Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

b. Transformer encoder

As a result of this, we have developed the Transformer encoder, which transforms the input token vectors into a sequence of context - dependent embedding vectors with shared form across many layers. In order to achieve quicker convergence, the pre-trained parameters from BERT [41] are employed.

c. Language modelling layers

In traditional BERT, it utilized only 15% masking for the next sentence prediction. It leads to suffering during the training process and consumes more computational time in the proposed approach. To train our PoS-BERT, we use the same settings and embedding as the ELECTRA training method. Generating $J_G(\theta)$ is the same as BERT training [41] in that it predicts token mask loss as well as the prediction of the masked next sentence.

To differentiate between tokens that have been replaced by generator G, discriminator D uses the generator G-predicted tokens, that is a naive binary classifier of each token with loss $J_D(\theta)$. Let's add together $J_G(\theta)$ and $J_D(\theta)$ to get the final language modelling loss $J_{lm}(\theta)$ by taking the output vector X of discriminator D and feeding it to these task-specific layers.

$$J_{lm}(\theta) = J_D(\theta) + J_G(\theta) \qquad (1)$$

d. PoS Layers

For linguistics tasks, we first reconstruct word representations from WordPiece tokens. The task-specific layers, such as the scoring layer and decoder layer, are then built. The word representations are built from the last token from the transformer. The scoring is performed by using a single layer feedforward network to train the POS tagging model and minimize the negative log-likelihood of the golden POS tag g_{pi} of every term (x_i), which would be represented as a cross entropy loss

$$J_t(\theta) = -logP_\theta(g_{pi}/x_i) \qquad (2)$$

e. Decoding Layer

Using the scoring layer and annotations, the decoder predicts the next sentence for the given input.

4.2 Hyper parameter tuning

Hyperparameter values are determined for the BERT model using a particle swarm technique. Optimization of particle swarms based on bird movements to find prey. Avian particles are utilised to identify the optimum hyperparameter for encoder, attention mechanism and hidden layer. Three limits are used here, ranging from [1 1 1 to 12 12 768]. Fifty repetitions were performed on each of the fifty particles or birds. In order to limit particle mobility, the L1 and L2 learning factors are employed, with values of 1.5 and 2. GES (Global Elite Solution) and LES (Local Elite Solution), are the two options for the speed upgrade. Speed updates are carried out using both GES and the local elite solution (LES) by the global elite solution (GES). As a result of these components, the next position (NP) and velocity update equation is provided in formula 3.

$$\begin{aligned} v[t+1] = v[t] + L1 * rand\ (1{,}0) & \qquad (3) \\ * (LES[t] - NP\ ) * L2 & \\ * rand(1{,}0) & \\ * (GES[t\ ] - NP[t\ ]) & \end{aligned}$$

$$NP[t = 1] = NP[t] + v[t]$$

It is possible to identify the appropriate hyperparameters of a BERT model for a downstream job using the techniques outlined above.

EVALATION

The accuracy and F1-score were used to evaluate the suggested method's performance. Accuracy shows that the testing dataset accurately identified the review type. The F1-score is based on the positive and negative predictive value, i.e., it reflects the ability to accurately identify both good and negative evaluations.

5. Implementation and Discussion

Using Python3 and CoLab, the proposed technique is implemented on Windows 10. First, the BERT model is pre-trained with PoS-masked words. The model's hyperparameters were tuned using a particle swarm optimization technique. A fine-tuned BERT model is applied to IMDB movie and hotel reviews [42] using the IMDB movie and hotel review databases in order to categorise reviews.

5.1 Movie review

A total of 50,000 movie reviews were found on the Kaggle dataset website. The evaluation contains both positive and negative aspects. The dataset is divided into two sections: training and testing. A total of 35,000 data points are used for training and another 15,000 are used for testing. The BERT fine-tuning approach uses parts of speech as a masking procedure and its noun masked texts as input tokens. The categorization task's result is the linked classes. Using particle swarm optimization methods, the optimum hyperparameters encoders, hidden size, and self-attentive heads were then identified for various input sizes, as shown in the table.

Table 1. Proposed method performance evaluation

Metric	Encoders				Self-attention (A)				Hidden layer				
	3	6	9	12	3	6	9	12	600	650	700	750	768
Accuracy using PSO	85	89.2	94.1	94.8	85	87	94.2	94.2	86	88.2	92	93.8	94.4
Accuracy Using PoS Masking-PSO	87	90.9	95	95.2	87	89	95.4	95.6	90	90.5	93.5	94.2	95.6
F1-score using PSO	81.2	84.6	94.5	94.5	81	84	93.5	94.5	82	86.2	91	92.5	94.5
F1-score Using PoS Masking-PSO	83.4	86.5	95.4	95.6	83.4	86	95.6	95.4	84	88	93.3	94.2	95.6

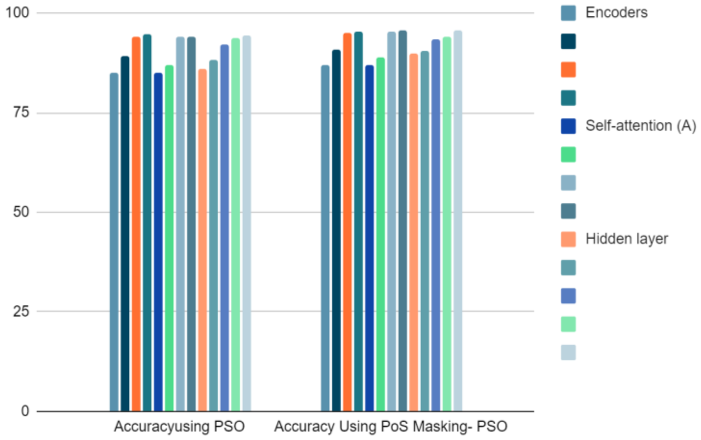


Figure 3. Accuracy comparison for movie review

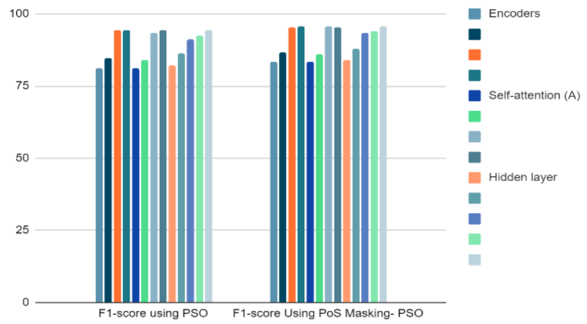


Figure 4. F1-score comparison for movie review

In comparison with basic hyperparameter tuning, parts of speech-based masking and particle-swarm based hyperparameter tuning increased prediction accuracy and F1-score for the movie review dataset It discovered that encoders with 9 attention mechanisms and 750 hidden layers functioned well for the proposed fine-tuning approach.

5.2 Hotel review dataset

This data may be seen on the Kaggle site [21]. On the Trip Advisor website, twenty thousand reviews were gathered from people who had been to the location. From 1 to 5 stars, users gave their opinions on this product. One-star reviews are known as negative reviews, whereas five-star reviews are considered positive reviews. Based on this, the dataset reorganised the data labels. According to the labelled data, the data splits into two categories of training and testing data. Seventy percent of the data utilized for training and thirty percent comes for testing, according to the statistics. Hotel review sentences were specified as input tokens for a parts-of-speech-based masking approach, and the results were used as training samples for the BERT fine-tuning procedure. As a consequence of the classification task, the matching class is returned. The Particle Swarm Intelligence technique is then used to choose the best hyperparameter encoder, concealed size, and self-attention head.

Table2. Proposed method evaluation on hotel review

Metric	Encoders				Self-attention (A)				Hidden layer				
	3	6	9	12	3	6	9	12	600	650	700	750	768
Accuracyusing PSO	90	95.3	95.4	95.6	91	95.6	95.8	95.0	91.1	95.2	95.6	95.8	95.9
Accuracy Using PoS Masking- PSO	91.6	96.7	96.6	96.6	92.2	96.7	96.8	96.5	92.3	96.4	96.6	96.8	96.8
F1-score using PSO	81.2	94.0	94.5	94.5	81	84	93.5	94.5	82	86.2	91	92.5	94.5
F1-score Using PoS Masking- PSO	83.4	86.5	95.4	95.6	83.4	86	95.6	95.4	84	88	93.3	94.2	95.6

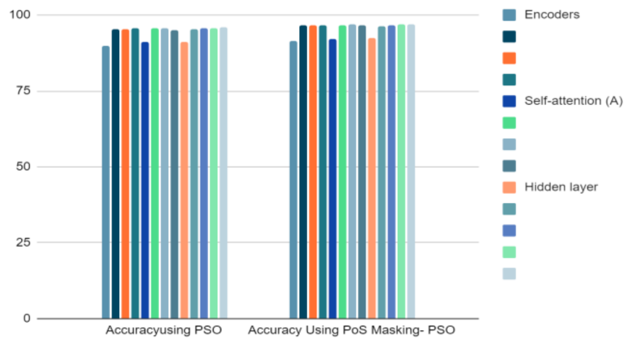


Figure5. Hotel Review accuracy comparison

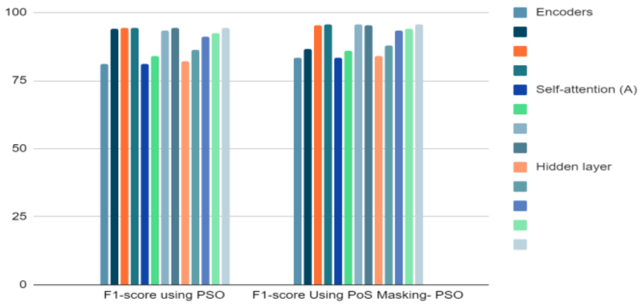


Figure6. Hotel review F1-score comparison

The tables and figure above show how part-of-speech masking and particle swarm hyperparameter adjustment improve prediction accuracy and F1 score on the Hotel review dataset. It was also observed that 6 encoders with 6 attention mechanisms and 750 or 768 hidden layers produced identical results. As a consequence, 6 encoders with 6 attention mechanisms and 750 hidden layers may be used for the hotel review dataset. A comparison of the proposed method performance with various existing approach is shown in the table 2.

Table 2. BERT Model performance comparison on testing set

Method	Accuracy		F1- score		Processing time (min)	
	Hotel review	Movie review	Hotel review	Movie review	Hotel review	Movie review
BERT	89	93.87	89.5	92.5	20	30
GS-BERT	89.5	91.5	89	91.2	25	40
GA-BERT	92	93.8	93.0	93.5	15	25
PSO-BERT	95.6	94.8	95	94.5	10	15
PoS Masking-PSO BERT	96.6	95.2	95.6	95.6	9	13

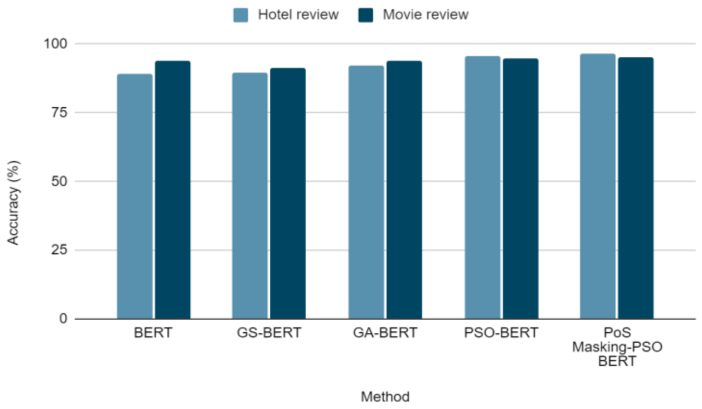


Figure7. Accuracy Comparison

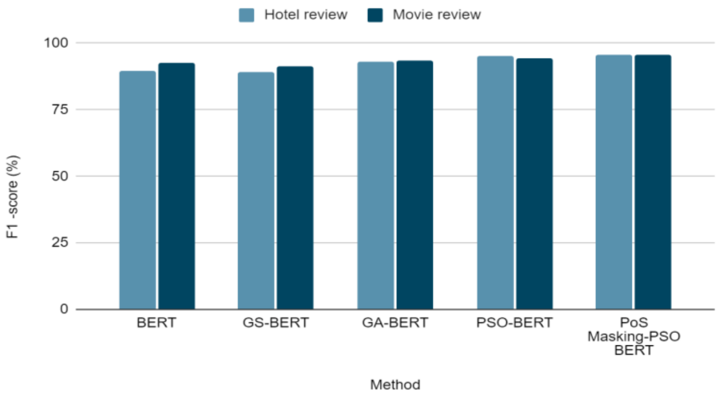


Figure7. F1-score comparison

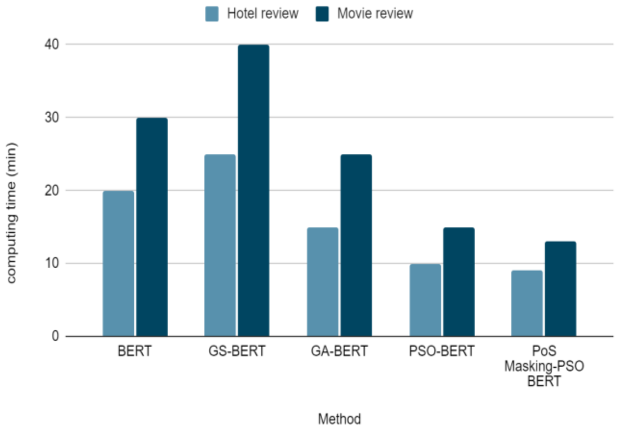


Figure8. Computing time comparison

The above performance comparison of different approaches highlighted the proposed parts of speech-based masking with particle swarm-based hyperparameter tuning in BERT to improve the classification performance on Review datasets like Hotel and Movie review databases as compared to the existing techniques. Therefore, the proposed approach is best for improving the BERT performance with minimal computation time for classification.

CONCLUSION

Literature survey looked at how to increase BERT's sentiment classification performance by modifying the BERT model for word embedding. The bulk of the research papers employed the concepts of pre-training and fine-tuning. Along with the fine-tuning procedure, hyper-parameter tweaking can enhance performance even further. In prior work, the hyper parameter tuning only performed so it required larger records for training. This problem is overcome with the proposed approach. Parts of speech-based masking in fine tuning and particle swarm based hyperparameter tuning in the BERT model improved the performance as compared to the single stage hyperparameter tuning performance. therefore, the proposed approach is best for improving the BERT performance and requires less processing time for testing process.

REFRENCES

[1] McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). Learned in translation: Contextualized word vectors. arXiv preprint arXiv:1708.00107.

[2] Howard, J., & Ruder, S. (2018). arXiv preprint arXiv:1801.06146.

[3] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

[4] Mohammad, S.M., 2021. Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text. In Emotion measurement (pp. 323-379). Woodhead Publishing.

[5] Banasik-Jemielniak, N.B.J.N. and Kałowski, P.K.P., 2022. Socio-cultural and individual factors in verbal irony use and understanding: What we know, what we don't know, what we want to know. Review of Communication Research, 10.

[6] Sabol, R. and Horák, A., 2022, September. New Language Identification and Sentiment Analysis Modules for Social Media Communication. In International Conference on Text, Speech, and Dialogue (pp. 89-101). Cham: Springer International Publishing.

[7] Zaki, M.Z., 2024. Revolutionising Translation Technology: A comparative Study Of Variant Transformer Models-BERT, GPT AND T5. Computer Science and Engineering–An International Journal, 14(3), pp.15-27.

[8] Wang, H., Li, J., Wu, H., Hovy, E. and Sun, Y., 2023. Pre-trained language models and their applications. Engineering, 25, pp.51-65.

[9] Selva Birunda, S. and Kanniga Devi, R., 2021. A review on word embedding techniques for text classification. Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020, pp.267-281.

[10] William, P., Shrivastava, A., Chauhan, P.S., Raja, M., Ojha, S.B. and Kumar, K., 2023. Natural Language processing implementation for sentiment analysis on tweets. In Mobile Radio Communications and 5G Networks: Proceedings of Third MRCN 2022 (pp. 317-327). Singapore: Springer Nature Singapore.

[11] Areshey, A. and Mathkour, H., 2024. Exploring transformer models for sentiment classification: A comparison of BERT, RoBERTa, ALBERT, DistilBERT, and XLNet. Expert Systems, 41(11), p.e13701.

[12] Sharma, N.A., Ali, A.S. and Kabir, M.A., 2024. A review of sentiment analysis: tasks, applications, and deep learning techniques. International journal of data science and analytics, pp.1-38.

[13] Briouya, A., Briouya, H. and Choukri, A., 2024. Overview of the progression of state-of-the-art language models. TELKOMNIKA (Telecommunication Computing Electronics and Control), 22(4), pp.897-909.

[14] Montrul, S., 2022. The psycholinguistics of heritage language acquisition. The Routledge handbook of second language acquisition and psycholinguistics, pp.73-84.

[15] Sun, C., Qiu, X., Xu, Y. and Huang, X., 2019, October. How to fine-tune bert for text classification?. In *China national conference on Chinese computational linguistics* (pp. 194-206). Cham: Springer International Publishing.

[16] Nassiri, A.K., Pernelle, N., Saïs, F. and Quercini, G., 2022. Knowledge Graph Refinement based on Triplet BERT-Networks. *arXiv preprint arXiv:2211.10460*.

[17] Mosbach, M., Andriushchenko, M. and Klakow, D., 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*.

[18] Li, F., Jin, Y., Liu, W., Rawat, B.P.S., Cai, P. and Yu, H., 2019. Fine-tuning bidirectional encoder representations from transformers (BERT)–based models on large-scale electronic health record notes: an empirical study. *JMIR medical informatics*, *7*(3), p.e14830.

[19] Ushio, A., Espinosa-Anke, L., Schockaert, S. and Camacho-Collados, J., 2021. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies?. *arXiv preprint arXiv:2105.04949*.

[20] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[21] Risch, J., Möller, T., Gutsch, J. and Pietsch, M., 2021. Semantic answer similarity for evaluating question answering models. *arXiv preprint arXiv:2108.06130*.

[22] Aoyama, T. and Schneider, N., 2022, July. Probe-less probing of BERT's layer-wise linguistic knowledge with masked word prediction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop* (pp. 195-201).

[23] Fu, L., Peng, H. and Liu, S., 2023. KG-MFEND: an efficient knowledge graph-based model for multi-domain fake news detection. *The Journal of Supercomputing*, *79*(16), pp.18417-18444.

[24] Srinivasagan, G. and Ostermann, S., 2024, June. HybridBERT-Making BERT Pretraining More Efficient Through Hybrid Mixture of Attention Mechanisms. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)* (pp. 285-291).

[25] Albu, I.A. and Spînu, S., 2022. Emotion detection from tweets using a BERT and SVM ensemble model. *arXiv preprint arXiv:2208.04547*.5zx [i]

[26] Tian, H., Gao, C., Xiao, X., Liu, H., He, B., Wu, H., Wang, H. and Wu, F., 2020. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. *arXiv preprint arXiv:2005.05635*.

[27] Yuan, J., Zhao, Y. and Qin, B., 2022. Learning to share by masking the non-shared for multi-domain sentiment classification. *International Journal of Machine Learning and Cybernetics*, *13*(9), pp.2711-2724.

[28] Zhang, X., van de Meent, J.W. and Wallace, B.C., 2021. Disentangling representations of text by masking transformers. *arXiv preprint arXiv:2104.07155*.

[29] Yang, D., Zhang, Z. and Zhao, H., 2022. Learning better masking for better language model pre-training. *arXiv preprint arXiv:2208.10806*.

[30] He, Z., Sun, T., Wang, K., Huang, X. and Qiu, X., 2022. Diffusionbert: Improving generative masked language models with diffusion models. *arXiv preprint arXiv:2211.15029*.

[31] Kesgin, H.T. and Amasyali, M.F., 2023, March. Iterative mask filling: An effective text augmentation method using masked language modeling. In *International Conference on Advanced Engineering, Technology and Applications* (pp. 450-463). Cham: Springer Nature Switzerland.

[32] Belfathi, A., Gallina, Y., Hernandez, N., Dufour, R. and Monceaux, L., 2024. Language Model Adaptation to Specialized Domains through Selective Masking based on Genre and Topical Characteristics. *arXiv preprint arXiv:2402.12036*.

[33] Liang, W. and Liang, Y., 2024. BPDec: Unveiling the Potential of Masked Language Modeling Decoder in BERT pretraining. *arXiv preprint arXiv:2401.15861*.

[34] Huang, C. and Shen, X., 2025, January. PoemBERT: A Dynamic Masking Content and Ratio Based Semantic Language Model For Chinese Poem Generation. In *Proceedings of the 31st International Conference on Computational Linguistics (pp. 50-60).*

[35] A survey of various embedding strategies used in NLP approaches was conducted by Liu et al .(2020) ELMO, GPT and its variations, BERT, XLNET, MASS, and BART are a few of the approaches. Each method was described, along with its analysis. It was mentioned that good hyper-parameter selection and pre-training in the future may enhance the effectiveness of NLP approaches.

[36] Mosbach, M, Andriushchenko, M & Klakow, D 2020, 'On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines', arXiv preprint arXiv:2006.04884.

[37] González-Carvajal, S. and Garrido-Merchán, E.C., 2020. Comparing BERT against traditional machine learning text classification. arXiv preprint arXiv:2005.13012.

[38] Yang, W, Zhang, H & Lin, J 2019, 'Simple applications of BERT for ad hoc document retrieva', arXiv preprint arXiv:1903.10972.

[39] Briskilal, J. and Subalalitha, C.N., 2022. An ensemble model for classifying idioms and literal texts using BERT and RoBERTa. Information Processing & Management, 59(1), p.102756.

[40] García Subies, G., 2021. Modelos de Transformers para la clasificación de texto (Doctoral dissertation, ETSI_Informatica).

[41] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019, June. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

[42] Hotel Reviews. (2021). Retrieved 18 June 2021, from https://www.kaggle.com/datafiniti/hotel-reviews

i