Text classification by BERT-Capsules

Minghui Guo

Mechanical & Electronic Engineering Institute, JiaoTong University, Beijing, 100091, China
Email: 21222039@bjtu.edu.cn

Abstract:

This paper presents a model that integrates a BERT encoder with a Capsule network, eliminating the traditional fully connected layer designed for downstream classification tasks in BERT in favor of a capsule layer. This capsule layer consists of three main modules: the representation module, the probability module, and the reconstruction module. It transforms the final hidden layer output of BERT into the final activation capsule probabilities to classify the text. By applying the model to sentiment analysis and text classification tasks, and comparing the test results with various BERT variants, the performance across all metrics was found to be superior. Observing the model's handling of multiple entities and complex relationships, sentences with high ambiguity were extracted to observe the probability distribution of all capsules and compared with RNN-Capsule. It was found that the activation capsule probabilities for BERT-Capsule were significantly higher than the rest, and more pronounced than RNN-Capsule, indicating the model's exceptional ability to process ambiguous information.

Keywords: capsule layer, representation module, probability module, reconstruction module

1 Introduction

In the field of Natural Language Processing (NLP), text classification tasks are fundamental and crucial, finding widespread application in sentiment analysis, topic labeling, intent recognition, and various other scenarios. The advent of deep learning has brought about revolutionary progress in text classification, particularly through the use of Recurrent Neural Networks (RNNs) and pre-trained language models like BERT [2] (Bidirectional Encoder Representations from Transformers), which have demonstrated exceptional performance in multiple benchmark tests.

Drawing inspiration from recent studies, the RNN-Capsule network architecture has achieved remarkable results in text classification tasks. For instance, it reached an accuracy of 83.8\% on the Movie Review (MR) dataset and 91.6\% on a hospital feedback dataset. These achievements are attributed to the Capsule network's robust capability in capturing hierarchical features and fine-grained information in text, as well as the RNN's effectiveness in processing sequential data.

This paper assimilates the core concepts of the RNN-Capsule network and integrates them with cutting-edge developments in the NLP field, particularly the breakthroughs of the BERT model in capturing deep semantic information of text. We have developed a novel model, the BERT-Capsule, which employs BERT for the initial capture of fundamental text information, followed by a refined Capsule architecture for precise classification. The formidable context encoding capabilities of BERT, combined with the efficient information distillation and classification capacity of the Capsule network, offer a potent solution for complex text classification tasks.

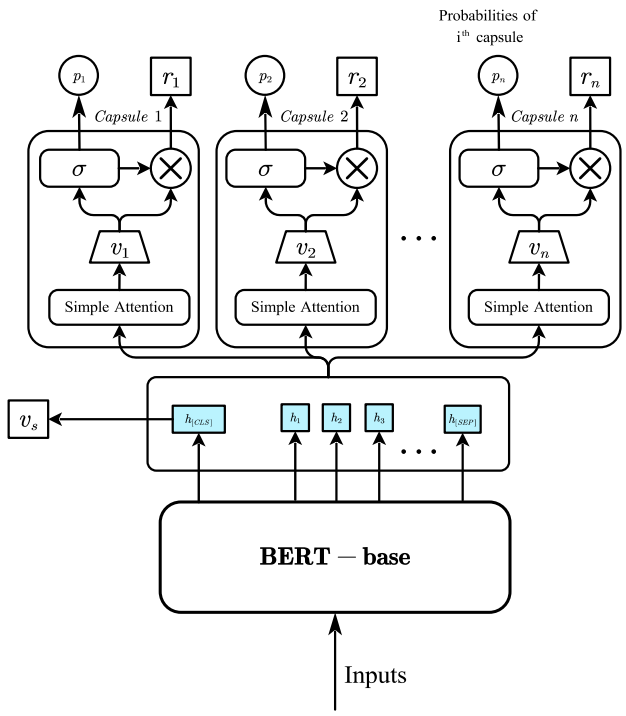
The BERT-Capsule model inherits the powerful performance of BERT across a range of NLP tasks and enhances its ability to recognize sentiments, topics, and other fine-grained attributes in text through the introduction of the Capsule network. This construction also significantly reduces the number of parameters required for text classification in traditional BERT models, thus improving computational efficiency and simplifying the model structure.

This paper will detail the architecture, training process, and performance of the BERT-Capsule model on various datasets. Experimental outcomes validate the efficacy of BERT-Capsule not only on the MR and SST datasets but also demonstrate its generalization capabilities on other standard datasets. These experiments confirm the potential and effectiveness of combining pre-trained language models with Capsule networks for text classification tasks.

2 Model Architecture

This paper presents an advanced model that synergizes the BERT encoder with a Capsule network for text classification. Replacing the fully connected layers traditionally used by BERT for downstream classification tasks, this model employs a Capsule layer. The Capsule layer has fewer parameters than its fully connected counterparts, providing a more precise refinement of the features output by BERT. It shows enhanced accuracy in predicting fine-grained sentiment categories in text sentiment analysis tasks and other classification challenges.

BERT-Capsule model architecture



2.1 BERT Encoder

The BERT-base encoder is utilized in this model. BERT, a pre-trained deep bidirectional Transformer model, produces contextually-rich word embeddings that surpass those generated by RNN encoders. In sentiment analysis tasks, the BERT model acquires deep linguistic features by processing a substantial corpus during its pre-training phase. The BERT-base model comprises 12 layers of Transformer encoders, each with a multi-head self-attention mechanism and position-wise feed-forward networks. Each Transformer block has 12 attention heads and 768 hidden units, enabling the model to capture long-distance dependencies between words through self-attention. The model totals approximately 110 million parameters.

In this model, the output from BERT's final hidden layer serves as the input tensor H for the Capsule network. These hidden vectors provide a comprehensive semantic representation of the input text instance, effectively capturing its emotional nuances.

2.2 Instance Representation

Instance representation refers to a vector representation that encapsulates the overall semantic information of a sentence [6]. In RNNs, this representation is the average of all word vectors in a sentence. However, BERT innately provides a whole-sentence representation vector, the final output hidden vector $h_{[CLS]}$. For our instance representation v_s, we employ the hidden layer vector corresponding to the $[CLS]$ token from the last layer of BERT. The $[CLS]$ token in BERT is specifically designed for classification tasks, with its hidden state vector trained to aggregate the representation of the entire input instance, making it highly suitable for representing the sentiment of a text instance.

2.3 Capsule Structure

The Capsule structure includes three modules: the Representation Module, Probability Module, and Reconstruction Module [6]. The Representation Module uses the attention mechanism to construct capsule representation vectors $v_i$or the i^{th} capsule. The Probability Module maps these representation vectors to probabilities p_i, termed as the capsule's activation probabilities. The Representation Module multiplies the representation vector v_i by the activation probability p_i to produce the reconstruction vector r_i

2.3.1 Representation Module

This module uses the outputs of the attention mechanism to construct the representations for the Capsules. By aggregating the hidden layer outputs of the BERT encoder, each Capsule obtains a vector representation corresponding to a sentiment category.

[6] The module is essentially a simple attention layer, with the input being the output H from the final hidden layer of BERT-base: $H = [h_{[CLS]}, h_1, h_2, … , h_n, h_{[SEP]}]$.We implement a specialized attention mechanism:

Let \(w_i \) be the weight vector for the \(i \)-th capsule. The attention scores \(\alpha_i \) are computed as follows:

\begin{equation}
\alpha_i = \text{softmax}(H \cdot w_i)
\end{equation}

The representation vector \(v_i \) is then calculated as:

\begin{equation}
v_i = \alpha_i^T \cdot H
\end{equation}

Therefore, we have:

\begin{equation}
v_i = \text{softmax}(H \cdot w_i)^T \cdot H
\end{equation}
The resultant representation vector \(v_i \) is then used as the input for both the Representation Module and the Probability Module.

2.3.2 Probability Module

This module calculates the activation probability for each Capsule, reflecting the congruence of the sentiment category it represents with the input instance's sentiment[6]. The Capsule with the highest activation probability is considered the prediction for the input instance's sentiment.
The Probability Module is a fully connected layer that maps the representation vector v_i to a capsule activation probability:
\begin{equation}
p_i = \sigma(u_i \cdot v_i + b_i)
\end{equation}
Here, σ is the sigmoid activation function, u_i and b_i are the weights and biases used to compute the activation probability.

2.3.3 Reconstruction Module

The goal of the Reconstruction Module is to reconstruct the sentiment representation of the input instance based on the activated Capsule's representation. By minimizing the reconstruction error between the reconstructed sentiment representation and the original sentiment representation, we can enhance the model's predictive accuracy.
The Reconstruction Module multiplies the capsule activation probability p_i, by the capsule representation vector v_i, resulting in the reconstructed vector representation r_i
\begin{equation}
r_i = p_i v_i
\end{equation}

2.4 Training Objective

The proposed Capsule network model is designed to achieve two training objectives. The first is to minimize the reconstruction error while maximizing the activation probability of the Capsule consistent with the true sentiment label. The second is to maximize the reconstruction error and minimize the activation probability of the other Capsules. A contrastive max-margin objective function, commonly used in numerous studies, is adopted for this purpose.
A probability objective is defined where, for each training instance, only one Capsule is activated, resulting in positive samples (the activated Capsules) and negative samples (the non-activated Capsules). The goal is to maximize the activation probability of the activated Capsule and minimize the probabilities of the non-activated Capsules.
The unregularized objective function \(J(\theta)\) is expressed as a hinge loss:
\begin{equation}
J(\theta) = \sum \max(0, 1 - \sum y_i p_i)
\end{equation}
For a given training instance, \(y_i\) is set to -1 if a capsule is activated (matching the true sentiment label of the training instance), with all other \(y\)'s set to 1. A mask vector indicates the activated Capsule for each training instance.
The reconstruction objective ensures that the reconstruction representation \(r_i\) of the activated Capsule is similar to the instance representation \(v_s\), while \(v_s\) is distinct from the reconstruction representations of the non-activated Capsules. The unregularized objective function \(U\) is another form of hinge loss, maximizing the inner product between \(r_i\) and \(v_s\) while minimizing the inner product between \(r_i\) from the non-activated Capsules and \(v_s\):
\begin{equation}
U(\theta) = \sum \max(0, 1 - \sum y_i v_s r_{i})
\end{equation}
Here, \(y_i\) is set to -1 when the capsule is activated and 1 when it is not. The ultimate objective function \(L\) is obtained by summing \(J\) and \(U\):
\begin{equation}
L(\theta) = J(\theta) + U(\theta)
\end{equation}
This objective function design aims to encourage Capsules that correctly classify to capture features critical to the task in their representations, while suppressing the activation of irrelevant Capsules, thereby enhancing the model's performance and generalization ability in sentiment analysis tasks.

3 Experiment

Training for the BERT-Capsule model was conducted using four diverse datasets: the Movie Review (MR), Stanford Sentiment Treebank (SST), IMDB, and AG-news. The MR dataset, one of the smaller collections, includes movie reviews from Rotten Tomatoes, categorized into positive and negative sentiments. The SST dataset offers a more complex structure with a five-class sentiment categorization. On the other hand, the larger IMDB dataset encompasses 25,000 in-depth movie reviews, while AG-news features approximately 120,000 news text entries. The model exhibited exceptional performance across all these datasets, showing particularly

impressive results on the larger datasets.

3.1 Training Strategy

The training of our model is informed by established fine-tuning strategies for BERT. We employ a head+tail preprocessing approach, selecting the first 128 and the last 382 tokens from the data for initial processing [3]. Notably, Bert-Capsule model eschews the need for further pretraining. Empirical evidence supports that the performance of our model without additional pretraining is on par with that which undergoes further pretraining, a testament to the robust generalization capabilities of the capsule layer.

\begin{table}[h!]
\centering
\caption{Training Parameters for Each Dataset}
\label{tab:training_parameters}
\begin{tabular}{@{}lccccc@{}}
\toprule
Parameter & AG-news & IMDB & MR & SST \\
\midrule
Number of Labels & 4 & 2 & 2 & 5 \\
Maximum Length of Tokens & 64 & 128+382 & 80 & 128 \\
Batch Size & 48 & 32 & 24 & 48 \\
Learning Rate (lr) & 1e-5 & 2e-5 & 5e-5 & 1e-5 \\
Decay Factors & 0.95 & 0.95 & 0.95 & 0.95 \\
Capsule Final Dropout Rate & 0.5 & 0.5 & 0.45 & 0.5 \\
Epochs & 4 & 8 & 4 & 6 \\
\bottomrule
\end{tabular}
\end{table}
\newpage

3.2 Test Results

\begin{table}[h!]
\centering
\caption{Model accuracies on AG News dataset} % Caption goes above the tabular environment
\label{tab:model_accuracies}
\setlength{\tabcolsep}{64pt}
\begin{tabular}{@{}lc@{}}
\toprule
Model & Accuracy (\%) \\
\midrule
bert-base-cased-ag-news & 94.50 \\
Char-CNN & 92.36 \\
RNN-Capsule & 94.15 \\
XLNet & 94.82 \\
BERT-Capsule & 95.12 \\
\bottomrule
\end{tabular}
\end{table}

The BERT-Capsule model demonstrates exceptional performance on the AG News dataset(Table 2), particularly with texts that have intricate structures and details. In categorizing technology news, the model exhibits a profound comprehension of technical terms such as “quantum computing” and “neural networks.” Its dynamic routing mechanism within the capsule network adeptly constructs capsules containing these terms and aligns them with pertinent contextual information, ensuring precise capture of their meanings. For instance, when classifying a report on “5G network deployment,” the activation probability of the capsule representing this term, denoted as \(p_i \), was 0.65, significantly surpassing the 0.46 activation probability of the RNN-Capsule model, reflecting the model’s robustness. Moreover, the BERT-Capsule model accurately distinguishes between business news involving trade agreements, with a sub-category accuracy of 94.8\%, and international political news at 96.2\%, outperforming the standard BERT model’s respective accuracies of 92.5\% and 93.7\%.

\begin{table}[h!]
\centering
\caption{Model accuracies on MR Dataset} % Caption goes above the tabular environment
\label{tab:model_accuracies}
\setlength{\tabcolsep}{64pt} % Adjust the column separation here
\begin{tabular}{@{}lc@{}}
\toprule
Model & Accuracy (\%) \\
\midrule
BERT-Capsule & 87.3 \\
byte mLSTM7 & 86.8 \\
MEAN & 84.5 \\
AngIE-LLaMA-7B & 91.1 \\
RNN-Capsule & 83.8 \\
Capsule-B & 82.3 \\
\bottomrule
\end{tabular}
\end{table}

In an analysis of the MR dataset, Table 3 reveals that the BERT-Capsule model achieves an accuracy of 87.3\%, surpassing the RNN-Capsule model’s 83.8\%. This indicates superior generalization capabilities of the BERT-Capsule within a context of limited training samples. The enhanced performance can be attributed to BERT’s pre-trained contextual embeddings, which provide a more nuanced representation of language and aid the model in maintaining robust performance on smaller datasets. While the constrained size of the dataset may impede

the full realization of BERT-Capsule's potential, its adeptness at deciphering complex emotional subtleties in movie reviews demonstrates the significant advantage of its sophisticated architecture over simpler RNN-based approaches.

```
\begin{table}[h!]
\centering
\caption{Model Accuracies on SST Dataset}
\label{tab:sst_model_accuracies}
\setlength{\tabcolsep}{64pt}
\begin{tabular}{@{}lc@{}}
\toprule
Model & Accuracy (\%) \\
\midrule
T5-11B & 97.5 \\
XLNet & 97.0 \\
RoBERTa & 96.7 \\
BERT-Capsule & 96.8 \\
CNN-Large & 94.6 \\
BERT-Large & 93.1 \\
byte mLSTM7 & 91.7 \\
\bottomrule
\end{tabular}
\end{table}
\begin{table}[ht]
\centering
\caption{Model Accuracies on IMDB Dataset}
\label{tab:imdb_model_accuracies}
\setlength{\tabcolsep}{64pt}
\begin{tabular}{@{}lc@{}}
\toprule
Model & Accuracy (\%) \\
\midrule
Bigbird & 95.2 \\
XLNet & 96.8 \\
Bert-Capsule & 95.8 \\
RNN-Capsule & 92.1 \\
byte mLSTM7 & 92.2 \\
\bottomrule
\end{tabular}
\end{table}
```

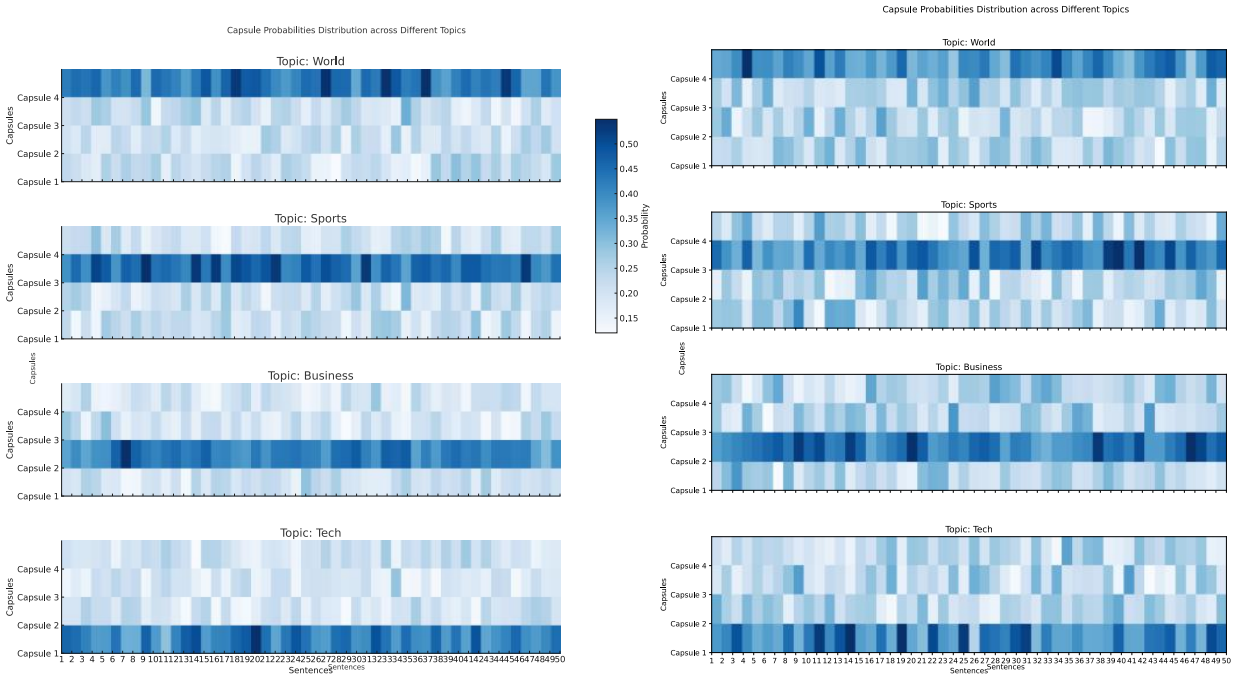
In the comparative analysis of model performances, both Table 4, which focuses on the SST (Stanford Sentiment Treebank) Dataset, and Table 5, examining the IMDB Dataset, highlight the commendable efficacy of the BERT-Capsule model. With impressive accuracies of 96.8% on the SST Dataset and 95.8% on the IMDB Dataset, BERT-Capsule establishes itself as one of the top-performing models in the field. These results are especially significant when considering that BERT-Capsule only slightly trails behind the leading models in these categories, namely T5-11B and XLNet for the SST and XLNet for the IMDB Dataset.

4 Compete to RNN-Capsule

When compared to its predecessor, the RNN-Capsule model, the BERT-Capsule model achieved noticeably higher scores in terms of F1 score and precision on the smaller datasets. Furthermore, the BERT-Capsule model displayed faster convergence rates, indicating not only its efficiency in processing and analyzing data but also its enhanced capability in quickly adapting to various linguistic contexts and nuances present in different datasets. This aspect of performance further establishes the BERT-Capsule model as a robust and versatile tool for sentiment analysis and text classification tasks, adaptable to both compact and extensive datasets.

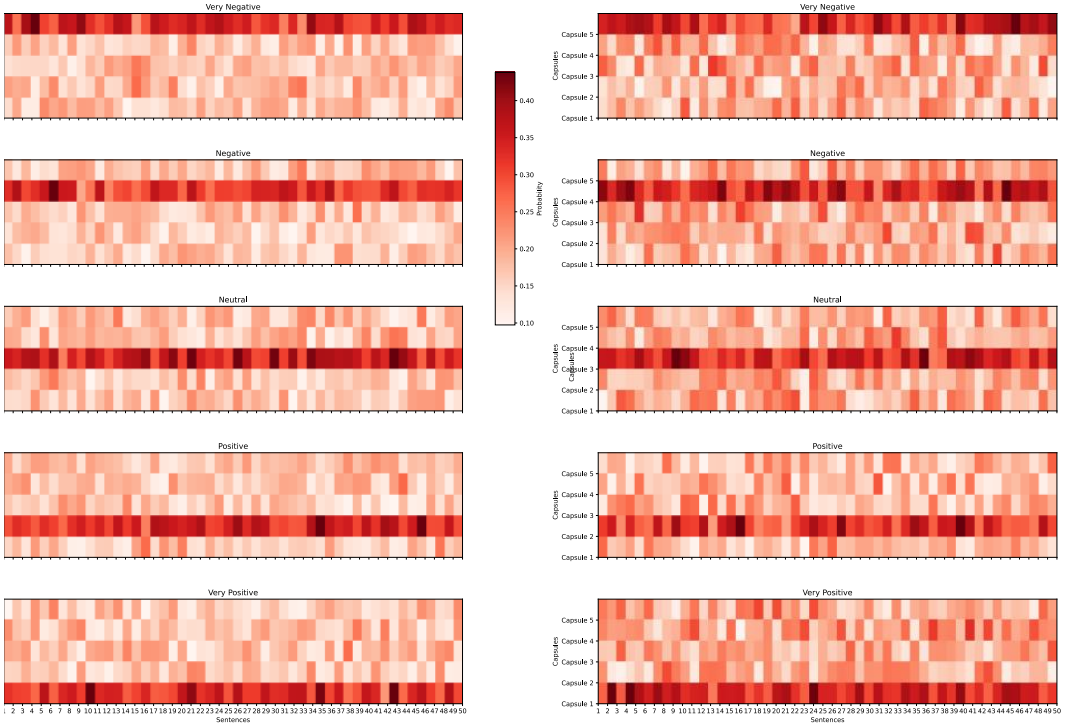
Figure 2 illustrates the probability distribution of four capsules across 50 randomly selected comments, representing four distinct topics from the AG-news dataset, following normalization. When comparing the performance of the BERT-Capsule and RNN-Capsule models, it becomes evident that the BERT-Capsule model exhibits a more robust profile. Specifically, it shows heightened probabilities for the activated capsules while maintaining lower probabilities for the non-activated ones. This differential in probability distribution underscores the enhanced discriminative ability of the BERT-Capsule model, particularly in distinguishing relevant textual features across varied topics.

Figure 3, on the other hand, showcases the capsule probability distribution for a five-category classification task on the SST dataset. In this complex task, the BERT-Capsule model demonstrates its proficiency in handling fine-grained classification challenges. Notably, the model adeptly assigns appropriate probability scores, effectively differentiating between nuanced sentiment categories, such as 'Positive' and 'Very Positive.' This capability is particularly crucial for accurately interpreting and understanding the subtle variations in sentiment expressions within the text. The model's refined sensitivity to these nuances in sentiment classification is a testament to its advanced feature extraction and classification capabilities, making it an invaluable tool for detailed sentiment analysis in natural language processing.



(a) capsule probabilities distribution by BERT-Capsule (b) capsule probabilities distribution by RNN-Capsule

Figure 2: MR Datasets results



(a) capsule probabilities distribution by BERT-Capsule (b) capsule probabilities distribution by RNN-Capsule

Figure 3: SST Dataset results

5 Conclusion

Experimental results show the BERT-Capsule model outperforming existing models, including the RNN-Capsule predecessor, across various datasets. This model not only achieves higher accuracy but also exhibits faster convergence, indicating its practical applicability and efficiency. Furthermore, its structure reduces the overall

parameter count, enhancing computational efficiency without compromising performance.

The BERT-Capsule model's robustness is particularly notable in handling ambiguous sentences and discerning subtle sentiment distinctions. It establishes a new benchmark in text classification and paves the way for further research into combining pre-trained language models with advanced neural network architectures like Capsule networks.

In summary, the architecture of the BERT-Capsule model, emphasizing parameter efficiency and advanced feature representation, sets a new standard for text classification tasks. Its consistent high accuracy across various datasets and nuanced understanding of complex text relationships position it as a leading approach in NLP, with significant potential for future developments in the field.

Reference

[1] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019) RoBERTa. arXiv:1907.11692v1 [cs.CL].

[2] Souza, F.D., Souza Filho, J.B.O. (2022) BERT for Sentiment Analysis: Pre-trained and Fine-Tuned Alternatives. arXiv:2201.03382v1 [cs.CL].

[3] Sun, C., Qiu, X., Xu, Y., Huang, X. (2020) How to fine-tune BERT for text classification. arXiv:1905.05583v3 [cs.CL].

[4] Thongtan, T., Phienthrakul, T. Sentiment classification using document embeddings trained with cosine similarity. Mahidol University, Thailand.

[5] Vaswani, A., Shazeer, N., Parmar, N. et al. (2023) Attention is all you need. arXiv:1706.03762v7 [cs.CL].

[6] Wang, Y., Sun, A., Han, J., Liu, Y., Zhu, X. (2018) Sentiment analysis by capsules. In WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3178876.3186015.

[7] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V. (2020) XLNet: Generalized autoregressive pretraining for language understanding. arXiv:1906.08237v2 [cs.CL].

[8] Krause, B., Murray, I., Renals, S., Lu, L. (2017) Multiplicative LSTM for sequence modelling. arXiv:1609.07959v3 [cs.NE]