Research on the application of deep learning-based BERT model in sentiment analysis

Yichao Wu[1a,5,*,†], Zhengyu Jin[1b,†], Chenxi Shi[2], Penghao Liang[3], Tong Zhan[4]

[1a]Computer Science, Northeastern University, Boston, MA, USA
[1b]Informatics, Univeristy of California, Irvine, Irvine, CA, USA
[2]Software development, Telecommunication Systems Management, Northeastern University, Boston, MA, USA
[3]Information Systems, Northeastern University, Boston, MA, USA
[4]Computer Science, Columbia University, NY, USA

[5]wu.yicha@northeastern.edu
*corresponding author
[†]These authors are co-first authors.

Abstract. With the rapid expansion of the Internet and social media, there has been an explosion of text data, containing valuable emotional insights. Sentiment analysis, aimed at discerning and processing people's emotional inclinations, opinions, and perspectives, has emerged as a pivotal area within natural language processing (NLP). This research explores the application of Deep Learning-based BERT models, particularly DistilBERT, in sentiment analysis. It highlights the importance of sentiment analysis in understanding emotional insights from vast text data and introduces BERT models as revolutionary methodologies for enhancing accuracy and efficiency. The study conducts experiments using the SST2 dataset, showcasing the effectiveness of DistilBERT in sentiment classification tasks. Overall, it underscores the transformative potential of BERT models in revolutionizing sentiment analysis methodologies and driving advancements in natural language processing research.

Keywords: Sentiment analysis, deep learning, BERT model, fine-tuning, experimental research.

1. Introduction

It finds applications across various domains including product review analysis, public opinion monitoring, and consumer behavior prediction, aiding companies and organizations in gaining deeper insights into public sentiment and facilitating more informed decision-making processes. Deep learning technology has revolutionized sentiment analysis research and applications by offering novel methodologies for handling complex text data. In contrast to traditional rule-based or machine learning approaches, deep learning techniques have the capability to autonomously learn features from extensive datasets, thereby significantly enhancing the accuracy and efficiency of sentiment classification and analysis. Among these methodologies, the BERT (Bidirectional Encoder Representations from Transformers) model stands out as a pioneering deep learning framework, leveraging pre-training to extract deep semantic information from text. It has achieved remarkable breakthroughs across various

[1] NLP tasks, including sentiment analysis, owing to its bidirectional Transformer architecture which enables a comprehensive understanding of contextual language relationships. However, despite its effectiveness in numerous domains, the optimal practices and efficacy of the BERT model for specific sentiment analysis tasks require further exploration and experimentation. This paper aims to address this gap by providing an overview of sentiment analysis fundamentals, tracing the advancements of deep learning in this domain, delving into the architecture and characteristics of the BERT model, and ultimately conducting experimental research to validate its application and optimization strategies in sentiment analysis.

2. Related work

2.1. Sentiment analysis review

With the development of social media and the increasing richness of information content and carriers, the construction of multimodal information such as text, vision and hearing provides a new opportunity for the development of sentiment analysis. Nowadays, more and more people like to record their feelings about the use of a product, the evaluation of a brand, or the opinion of a social event through video or audio, and share it on social media. Compared with text, video has great potential for diversity in content creation, and is often more attractive and convincing, so its influence and radiation are also wider. In such video or audio, the facial expressions of the characters, the voice intonation of the human voice, the melody of the background music, and even physiological signals such as brain electricity are all effective tools to convey emotional information. Therefore, in recent years, multimodal sentiment analysis has become a hot research direction in the field of natural language processing and computer vision. It can be predicted that in the future for a long time, this [2] multimodal sentiment analysis method will become the main means of multimedia content analysis and understanding, and can be widely used in a variety of scenarios, such as stock market prediction, election polls, customer satisfaction assessment, news public opinion perception, advertising recommendation, etc.

2.2. Application of deep learning to NLP

Deep learning has become the dominant method of natural language processing (NLP) research, especially in large-scale corpora. In natural language processing tasks, sentences are often thought of as a series of markers. Therefore, popular deep learning techniques such as recurrent neural networks (RNN) and convolutional neural networks (CNN) have been widely used in text sequence modeling. As a result, these graphically structured data can encode complex paired relationships between entity tags to learn more information representations. However, deep learning techniques are known to be disruptive to Euclidean data (such as images) or sequential data (such as text), but are not immediately applicable to graph-structured data. As a result, this gap has driven research into deep learning for graphs, particularly the development of graph neural networks (GNN). There has been a surge of interest in applying/developing various types of GNN and considerable success in many natural language processing tasks, from classification tasks such as sentence classification, semantic role labeling, and relationship extraction, to generation tasks such as machine translation, question generation, and summarization.

2.3. Bidirectional Encoder Representations from Transformer

The full name of BERT model is: Bidirectional Encoder Representations from Transformer. As can be seen from the name, the goal of the BERT model is to use large-scale unlabeled corpus to train and obtain a Representation of a text containing rich semantic information, that is, a semantic representation of the text, and then fine-tune the semantic representation of the text in a specific NLP task, and finally apply it to that NLP task. Among them, the Attention mechanism is also the key part of the analysis. In essence, Attention is to select a small amount of important information from a large amount of information, and focus on these important information, ignoring most of the unimportant information.

The first process can be subdivided into two stages: the first stage calculates the similarity or correlation between Query and Key; In the second stage, the original score of the first stage is normalized. In this way, the Attention calculation process can be abstracted into the three stages shown in the figure.

3. BERT model foundation and methodology

3.1. Model architecture

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained natural language processing model released by Google in 2018. At the heart of the BERT model is the Transformer encoder, which allows unsupervised pre-training on a large corpus and then fine-tuning on a variety of NLP tasks through fine-tuning. BERT model is a bidirectional deep learning model that considers all words in context at the same time to better understand the meaning of a sentence. [3] BERT models have been shown to achieve the best results on a number of NLP tasks, including question answering, text classification, and named entity recognition.

- BERT is a natural language understanding model based on deep neural networks that can learn the semantics and structure of a language from large-scale unlabeled text.
- BERT's innovation lies in its use of a bi-directional Transformer encoder that can take into account context information in both left and right directions to capture richer language features.
- BERT's pre-training tasks include MLM and NSP, which are used to learn vocabulary and sentence-level representations, respectively.
- BERT not only improves the performance of the model, but also simplifies the fine-tuning process of the model, which can be adapted to different tasks by adding only a small number of task-related layers on top of the pre-trained model.
- These models optimize or innovate BERT in different ways, such as increasing the amount of data, reducing the number of parameters, and changing the pre-training task.

3.2. Pre-training task

In BERT, language models are built in two pre-trained ways:

1) BERT language model Task 1: MASKED LM

In BERT, Masked LM (Masked language Model) built a language model, which was also one of BERT's pre-training tasks. Simply put, masked or replaced any word or word in a sentence randomly. Then, the model is asked to predict the covered or replaced part through understanding the context, and then only calculate the Loss of the covered part when making Loss, which is actually an easy task to understand, and the actual operation is as follows:

1. These tokens have an 80 chance of being replaced with mask;

2. There is a 10% chance that it will be replaced with any other token;

3. There's a 10% chance it's intact.

Then let the model predict and restore the hidden or replaced part, the final output of the hidden layer of the model's calculation results are:

$$X_{hidden} : batch\_size,\ seq\ len,\ embedding\_dim \tag{1}$$

This step requires initializing a mapping layer weight W_{vocab}:

$$Wvocab : [embedding\_dim,\ vocab\_size] \tag{2}$$

In use Wocab to complete the mapping of hidden dimensions to the number of word vectors, requiring only the matrix multiplication (dot product) of X_{xdden} and W_{vocab} :

$$X_{hidden}W_{vocab}: [batchsize,\ segien] \tag{3}$$

After vocabsize, the above calculation results are normalized in theVocab_size(the last) dimension so.ftmax is the sum of vocab_size corresponding to each word to 1. Then we can obtain the prediction results of the model through the word with the highest probability in V_{cocab}_size, and we can compare it with the prepared La bel makes a Loss and backpasses the gradient.

3.3. Adaptation and optimization of BERT model

In the process of realizing the adaptive optimization of BERT model, the following methods need to be adopted:

1) Improve the hidden language model

In BERT model, text preprocessing is segmented according to the smallest unit. For example, the pre-processing of English text uses Google's wordpiece method to solve the problem of unknown words. In the BERT-WWM model subsequently released by Google, the way of full word coverage is proposed. [5] BERT-Chinese-wwm uses Chinese word segmentation to cover all the words that make up a complete word at the same time. SpanBERT adopted geometric distribution to randomly sample the masked phrase fragments, and used Span boundary word vectors to predict the masked words

2)Noise reduction autoencoder is introduced

MLM randomly replaces the words in the original text with [MASK] tags, which itself destroys the text, equivalent to adding noise to the text, and then trains the language model to restore the text and eliminate the noise. DAE is an autoencoder with noise reduction function, which is designed to restore the input data containing noise to clean raw data. For the language model, it is to add the noise data to the original language, and then remove the noise through the model learning to restore the original text.

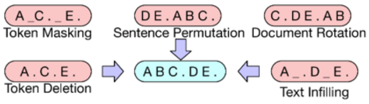

Figure 1. Model training sequence

3) Introducing alternative word detection

MLM makes predictions about the words marked by [MASK] in the text in an attempt to recover the original text. Its prediction may be completely correct, or it may predict a word that is not part of the original text.

ELECTRA introduces alternative word detection to predict which words in a sentence generated by a language model are words in the original sentence and which words are words generated by the language model and do not belong to the original sentence. ELECTRA uses a small [6] MLM model as a Generator to make predictions about sentences containing [MASK]. In addition, a Discriminator based on binary classification is trained to judge the sentences generated by the generator.

4. Experiment and conclusion

In the experiment, we will implement sentiment analysis methods using deep learning combined with BERT models and compare them with other traditional word vector models, including FastText, Word2Vec, and GloVe. In order to achieve the goal of the experiment, we will import each word vector model in turn (including FastText, Word2Vec, GloVe, and DistilBERT) and show how to load the pre-trained model and apply it to the sentiment analysis task.

4.1. Data set

The SST2 dataset was chosen as the dataset for the sentiment classification task in this paper mainly because it has the following advantages. First, the SST2 dataset contains a large number of film review texts from different channels and platforms, covering multiple linguistic styles and thematic content, thus enabling the model to learn a wider range of linguistic expressions and emotional experiences. In contrast, the advantages of choosing SST2 dataset in this paper lie in its rich text data and accurate emotion labels, which provide sufficient training and evaluation basis for deep learning models. Compared to datasets such as IEMOCAP and MELD, which are also commonly used in the field of emotion recognition, the SST2 dataset is more suitable for the emotion classification task in this paper because it has wider text coverage, more concise and clear emotion labels, and higher quality labeling, which helps to ensure that the model is accurately guided in training and evaluation. The performance and generalization ability of the model are improved. This dataset has the following characteristics:

Table 1. The sentiment analysis table uses the label 0/1 to represent positive and negative emotions

sentence	labels
a stirring, funny and finally transporting re imagining of beauty and the beast and 1930s horror films	1
apparently reassembled from the cutting room floor of any given daytime soap	0
They presume their audience won't sit still for a sociology lesson	0
this is a visually stunning rumination on love , memory , history and the war between art and commerce	1
jonathan parker 's bartleby should have been the be all end all of the modern office anomie films	1

SST2 dataset is widely used in the training and evaluation of sentiment analysis tasks, and is favored by researchers because it contains rich text data and accurate sentiment labels. In this experiment, we will use the SST2 dataset as training and test data to evaluate the performance of the proposed deep learning model on the emotion classification task.

4.2. Model building

The emotion classification model of sentences consists of two parts:

DistilBERT processes the input sentence and passes on some of the information it extracts from the sentence to the next model. DistilBERT is a smaller version of the BERT model that was open-source by the HuggingFace team. It retains BERT's power while being smaller and faster than BERT.A basic Logistic Regression model that will process the output of DistilBERT[7] and categorize the sentences, output 0 or 1.The data passed between the two models is a 768 dimensional vector.Assuming the sentence length is n, then a sentence passing through BERT should yield n 768 dimensional vectors.

Model training and prediction

1) Training

Two models are used in this article, but only the Logistic Regression model needs to be trained during the implementation.For the DistilBERT model, the parameters that were pre-trained by the model were used, and the model was not used to train and fine-tune the sentence classification task.Use the Scikit Learn toolkit to do so. Divide the entire BERT output into train/test datasets.

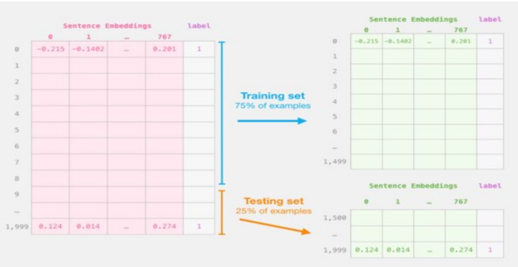

Figure 2. Step #2: Test/Train Split for model #2, logistic regression

The data is divided into a training set comprising 75% of the data and a test set containing the remaining 25%. Utilizing sklearn's train/test split function, the samples are shuffled before being split to ensure randomness in the selection process. Following this, the regression model is trained using various machine learning methods.And This table provides a comparison of the main hyperparameters used in your article for the DistilBERT and Logistic Regression models. The DistilBERT model has a learning rate of 5e-5, a batch size of 32, utilizes the AdamW optimizer during training, and applies a dropout rate of 0.1. On the other hand, the Logistic Regression model has a learning rate of 0.01, a batch size of 64, and a dropout rate of 0.5, without a specific optimizer mentioned during training.

4.3. forecast

In this experiment, we compared sentiment analysis methods using deep learning with BERT models against traditional word vector models like FastText, Word2Vec, and GloVe. We chose DistilBERT for text sentiment classification and evaluated its performance on the SST2 dataset. The model architecture included DistilBERT for processing input sentences and logistic regression for categorizing them. [8]split the data, trained the logistic regression model, and made predictions using tokenization and embedding mapping.

4.4. experimental results

BERT model training was carried out on the original data set. The output tensor size of BERT model was [2000, 59, 768]. # Vector corresponding to [CLS]token was extracted, and the results were as follows:

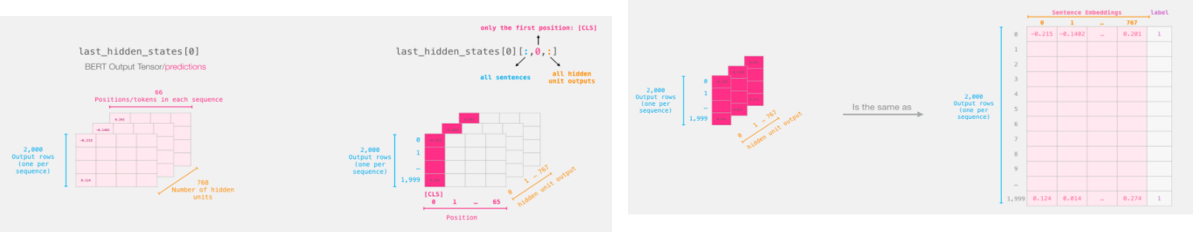

Figure 3. The tensor size of the BERT model output is [2000, 59, 768], the vector corresponding to the token

According to the above model, it can be found that the results of the trained classifier are obviously better than those of the random prediction.In the original tutorial, the accuracy of 91.3 DistilBERT was reached by fine-tuned Distilbert (see huggingface website for the model), and the BERT model with the full number of parameters can achieve a score of 92.7.

In the analysis of the experimental results, a deeper examination of the performance of the BERT model is warranted. This entails comparing its performance against expected outcomes, identifying potential areas for improvement, and discussing its practical implications for sentiment analysis tasks.comparing the achieved results with the expected outcomes provides insights into the effectiveness of the BERT model. [9-10] A comprehensive analysis of the BERT model's performance involves evaluating its performance against expectations, identifying areas for improvement, and discussing its practical implications for sentiment analysis tasks. This holistic approach facilitates a deeper understanding of the model's efficacy and informs future research directions in sentiment analysis.

5. Conclusion

Based on the research conducted on the application of Deep Learning-based BERT models in sentiment analysis, several key conclusions can be drawn：

Firstly, the study highlights the significant potential of incorporating BERT models into sentiment analysis tasks. The experiment compared the performance of DistilBERT with traditional word vector models like FastText, Word2Vec, and GloVe. Despite the lack of fine-tuning, DistilBERT showcased promising results, outperforming traditional models. This underscores the effectiveness of leveraging deep learning techniques, particularly BERT models, in enhancing sentiment analysis accuracy and efficiency. By fine-tuning the DistilBERT model, further enhancements in performance were observed, as evidenced by increased accuracy in downstream tasks. This emphasizes the significance of optimizing BERT models for specific sentiment analysis tasks, highlighting the potential for even greater improvements with tailored fine-tuning approaches.BERT's bidirectional Transformer architecture, more accurate and efficient sentiment classification can be achieved. Thus, the study underscores the transformative potential of BERT models in revolutionizing sentiment analysis methodologies and driving advancements in natural language processing research.

References

[1] Devlin J,Chang M W,Lee K,et al.BERT: pretraining ofdeep bidirectional transformers for langu age understandingl .arXiv e-print,2018,arXiv: 1810.04805

[2] Li Z Y,Zou Y C,Zhang C,et al.Learning implicit senti-ment in aspect¬based sentiment analysis with supervisedcontrastive pre-training [C]//Proceedings of the 2021Conference on Empirica l Methods in Natural LanguageProcessing.Online and Punta Cana,Dominican Republic.Strou dsburg, PA,USA:Association for ComputationalLinguistics,2021

[3] Klinger R, de Clercq O, Mohammad S M, et al.IEST:WASSA-2018 implicit emotions shared tas k [J].arXive print,2018,arXiv: 1809.01083

[4] Balazs J, Marrese-Taylor E,Matsuo Y. llIDYT at IEST2018: implicit emotion classification with deep contextu-alized word representations [C]//Proceedings of the 9thWorkshop on Comput ational Approaches to SubjectivitySentiment and Social Media Analysis.Brussels , Belgium. Stroudsburg,PA,USA:Association for ComputationalLinguistics,2018:50-56

[5] Chronopoulou A,Margatina A, Baziotis C,et al.NTUA-SLP at lEST 2018: ensemble of neural tr ansfer methodsfor implicit emotion classification [C]//Proceedings ofthe 9th Workshop on C omputational Approaches to Sub-jectivity,Sentiment and Social Media Analysis. Brussels,Be lgium.Stroudsburg,PA,USA: Association for Computa-tional Linguistics,2018:57-64.

[6] K. Tan and W. Li, "Imaging and Parameter Estimating for Fast Moving Targets in Airborne SA R," in IEEE Transactions on Computational Imaging, vol. 3, no. 1, pp. 126-140, March 2017, doi: 10.1109/TCI.2016.2634421.

[7] Srikanth Ryali, et al. "Deep learning models reveal replicable, generalizable, and behaviorally re levant sex differences in human functional brain organization.." Proceedings of the National Academy of Sciences of the United States of America 121. 9 (2024):

[8] K. Tan and W. Li, "A novel moving parameter estimation approach offast moving targets based on phase extraction," 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 2015, pp. 2075-2079, doi: 10.1109/ICIP.2015.7351166.

[9] Yuan Yuan, et al. "Supercritical carbon dioxide critical flow model based on deep learning." Progress in Nuclear Energy 170. (2024):

[10] Mengsheng Wang, et al. "Design of smart home system speech emotion recognition model based on ensemble deep learning and feature fusion." Applied Acoustics 218. (2024):