# A BERT-CNN Based Approach on Movie Review Sentiment Analysis

Bowen Zhang[1,*]

[1]Faculty of Engineering and Information Technology, University of Melbourne, Melbourne, 3010, Australia

**Abstract:** Sentiment analysis plays a vital role in the decision-making of multiple fields. Specifically, in movies and television, audiences' reviews can help with casting, the direction of the plot, etc. To further improve the performance of the original BERT model, a BERT-CNN-based approach is proposed in this paper to do sentiment analysis on IMDb dataset. Although their performances are nearly the same throughout the research, the BERT-CNN approach is better at negative sentiment detection. It got an average elevation of 3.6% in accuracy after the ensemble. Apart from that, topic modeling is also performed to show that most negative reviews are commented on from multiple aspects instead of criticizing only, making sentiment analysis of movie reviews a complex problem.

## 1. INTRODUCTION

Sentiment analysis, or opinion mining, is analyzing, processing, concluding, and reasoning subjective texts with personal emotions. The study can be considered a classification task, where a reader is generally classified as positive, negative, or neutral. With the rapid development of social media, sentiment analysis is required in multiple fields to collect public opinion for decision-making. A study by Joyce and Deng collected over 7.5 million opinion tweets about the two presidential candidates during the 2016 US Presidential election. Comparing sentiment expressed by these tweets with the polling data, they found a 94% correlation between [1]. Besides, business owners can use sentiment analysis in commodities and movies to deeper understand customers' attitudes and make improvements accordingly [2].

In this research, I intend to propose a deep learning model in a combination of BERT and CNN to do sentiment analysis on movie reviews and examine its performance compared with the sole BERT model. Apart from that, I explored topics of both positive and negative reviews to analyze user demands.

## 2. LITERATURE REVIEW

TF-IDF gives the relevance of a term to a document by multiplying the term frequency by its inverse document frequency. TF-IDF lowers the weight of stop words like articles or prepositions but gives high importance to familiar words in a small set of documents [3]. An example of this in sentiment analysis is the study by Das and Chakraborty, where they used TF-IDF to convert words into scaled vectors to train an SVM model. Compared with a binary BOW model, the performance of the TF-IDF is much higher. They also found that the model works better by combining subsequent word negation (NWN, negating the word after 'not') with TF-IDF [4]. With the deepening of the study on deep learning, most current studies focus more on various artificial neural networks. Rhanoui et al., proposed a framework for document-level sentiment analysis. They used Doc2vec to embed the documents as input and put them through a CNN plus BiLSTM framework. According to their research, the model outperformed others in newspaper article prediction with 90.66% accuracy [5]. Sousa et al., tested the performance of BERT on sentiment analysis and found that BERT outperforms all classical machine learning models like Naive Bayes and Support Vector Machines. After sentiment analysis on stock articles, they also provided relevant information for decision-making [6]. Dong et al., combined BERT with CNN to extract both local (CNN) and global (BERT) semantic features when doing sentiment analysis on commodity reviews and got an elevation on F1 score in comparison with pure BERT and pure CNN models [7].

## 3. DATASET

The dataset used in this research is about movie reviews from IMDb, an online database of information related to movies, and was first collected by Maas et al. This dataset consists of 50K studies with their corresponding sentiment labels split evenly into 25K training sets and 25K testing sets. The sentiment is judged based on the scores provided and the reviews so that a score ≥ 7 is considered positive, ≤ 4 is unfavorable, and those more neutral are excluded [8]. The number of positive and negative instances in training and testing sets are all 12,500 each. However, specific examples are duplicated during data pre-processing and appear in training and testing sets. This

[*]2049027754@qq.com

will be discussed in the next section.

## 4. METHODOLOGY

Three methods are implemented: logistic regression as a baseline model and two BERT-based models for performance comparison.

### 4.1 DATA PRE-PROCESSING

The data is given in .txt format. When loading data and forming the dataset, some movie reviews appear duplicated and occur in training and testing sets. The label distribution after removing these instances is shown in Figure. Further detailed pre-processing is discussed in the methodology section, but for all methods, the training data is split into 85% training and 15% parameter tuning.

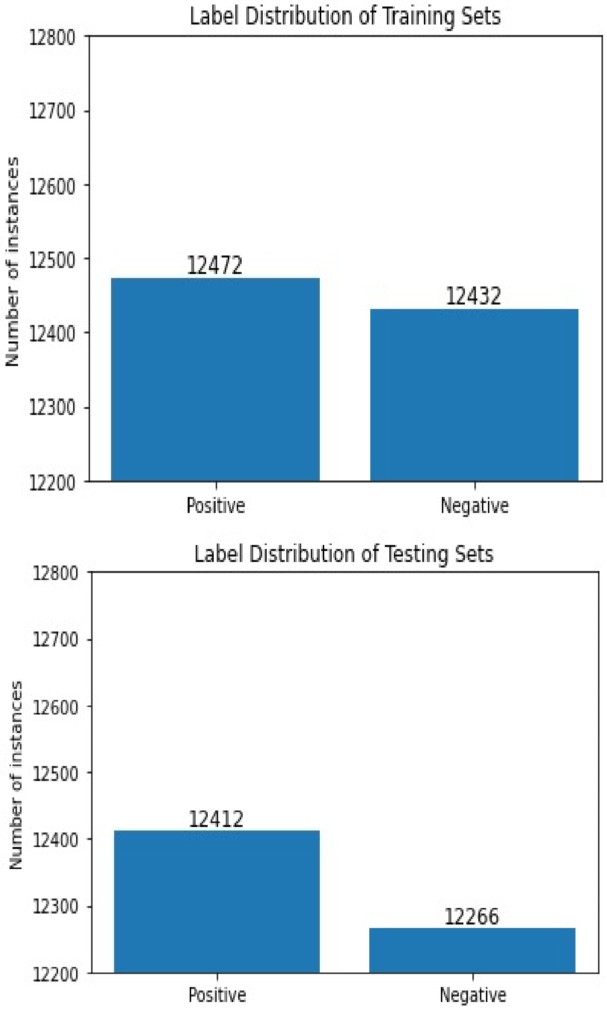


**Figure 1:** Label Distribution of the Dataset

### 4.2 Logistic REGRESSION

Logistic Regression is a supervised probabilistic approach, and its algorithm is considered a baseline for classification in natural language processing (NLP) [9]. A logistic regression model uses a logistic (sigmoid) function, shown in Formula 1, to form a valid probability when predicting $P(Y = 1|x)$, so an instance is classified as one if this value is more significant than 0.5 and 0; otherwise, in binary classification. Logistic Regression can also be applied in multinomial type by using the SoftMax function.

$$P(Y = 1|x) = \frac{1}{1+\exp(-x\prime w)} \quad (1)$$

The core idea of logistic regression is to give different features a proper weight, so after removing stop-words, each instance is processed into a unigram bag-of-words form. The regularization parameter is tuned on the validation set for optimal performance.

### 4.3 BERT

Bidirectional Encoder Representation from Transformers (BERT) is a language model proposed by Google AI. The architecture of a BERT model is a multi-layer bidirectional transformer encoder, and it learns to predict the characteristic representation of the input sequence (contextualized embedding). Since the architecture is bidirectional, the information from both sides is integrated so BERT can deal with various tasks [10].

The pre-trained BERT model used in this research is best-base-uncased. For the text to be suitable for BERT, a piece of review is prepended with a [CLS] token (classification), and a [SEP] token (separator) is added at the end. Because of the nature of BERT, which requires fixed-length sentences as inputs, a max length parameter is chosen before training so that sentences longer than this length are truncated, and those shorter are padded with [PAD] token [11]. This parameter is chosen based on the quantiles of the size after tokenization, and a boxplot is shown in Figure 2. Besides, the attention mask is also obtained (0 for [PAD] tokens and 1 for others) for the model to focus on non-padding tickets.

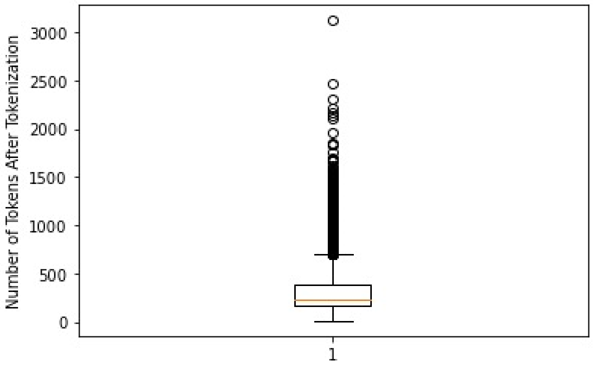


**Figure 2**: Number of Tokens After Tokenization

After passing into the BERT model, the embedding of the [CLS] token is extracted into a linear layer for classification, as it is where the contextualized representation of the movie review stores. The whole architecture is shown in Figure 3.

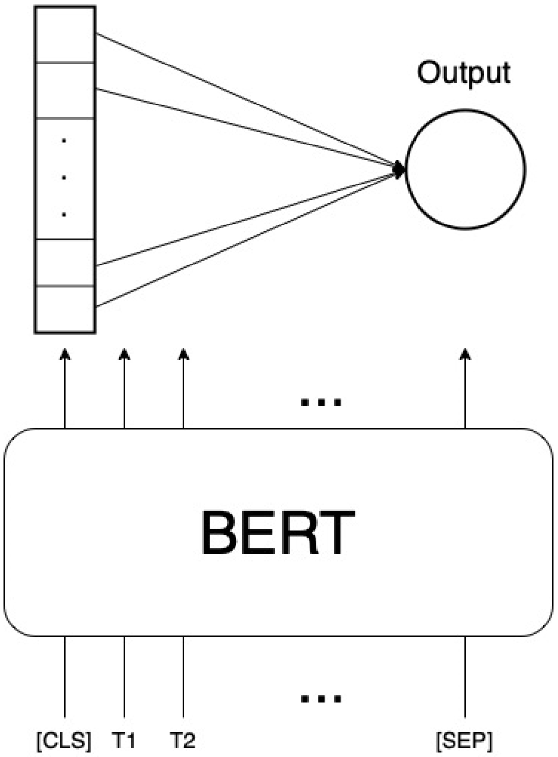


**Figure 3**: Sketch Map of BERT-Based Model

### 4.4 BERT-CNN

Convolutional Neural Network (CNN) is a feedforward neural network that utilizes convolution calculation. It is initially applied in computer vision to capture local features. In the NLP field, word vectors (embeddings) are fed into a 1D CNN instead of image pixels, and CNN can automatically capture and identify essential features very accurately. In this approach, apart from the [CLS] embedding, other words' embeddings are passed through a CNN to get a shorter but more concentrated representation. All the generated models are concatenated with the [CLS] embedding to be passed into a linear layer for classification. The contextualized information and the review's essential details can also be considered when identifying. The detailed framework is shown in Figure 4.

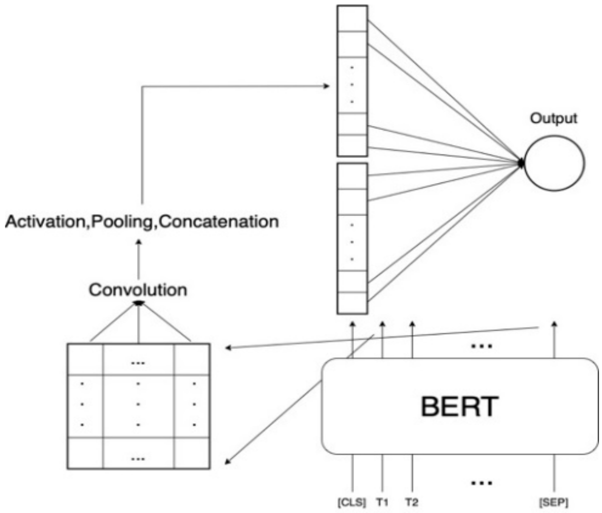


**Figure 4**: Sketch Map of BERT-CNN-Based Model

## 5. RESULT

The model performances and results on test sets are shown in Figure 5 and Table 1. It clearly shows that their performances almost converge after running for just one epoch, showing BERT's robustness.

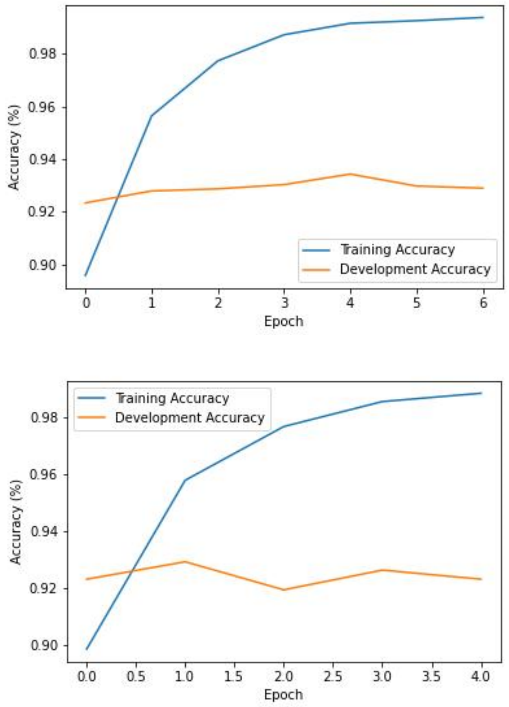


**Figure 5:** Model Performance of BERT (left) and BERT-CNN (right)

Table 1: Results of the Models (positive on the left; negative on the right)

| Model | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|
| Logistic Regression | 0.879\|0.874 | 0.875\|0.878 | 0.877\|0.876 | 0.876 |
| BERT | 0.942\|0.919 | 0.918\|0.942 | 0.930\|0.931 | 0.930 |
| BERT-CNN | 0.917\|0.942 | 0.944\|0.914 | 0.930\|0.927 | 0.929 |
| Ensemble | | | | 0.944 |

According to Table 1, the one with CNN does not surpass the pure BERT model. Still, it shows an interesting pattern that BERT is good at identifying positive sentiment and BERT-CNN is good at identifying negative emotion, so by the ensemble of the three models above, a moderately high accuracy of 94.4% is obtained.

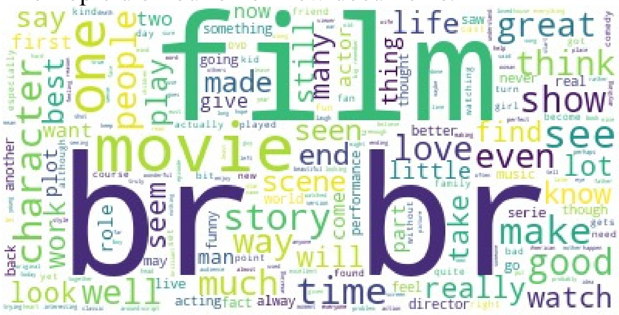
By analyzing those wrongly classified, 32.3% are longer than 400 after tokenization, so that some critical information may be missed during processing. Also, by scanning through some of these reviews, there is a clear trend that the wording of these reviews is not very extreme. They may criticize some scenes but still give a high mark, which is considered positive sentiment in this dataset.

## 6. TOPIC MODELING

Topic modeling is done in this section to detect the reason behind user ratings automatically. A topic model is an unsupervised statistic model to learn common, overlapping themes in a series of documents.

### 6.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is an algorithm of the topic model. It is a Bayesian model with both prior and posterior Dirichlet distribution. Compared with probabilistic latent semantic analysis (PLSA), LDA can infer topic distribution on new documents.



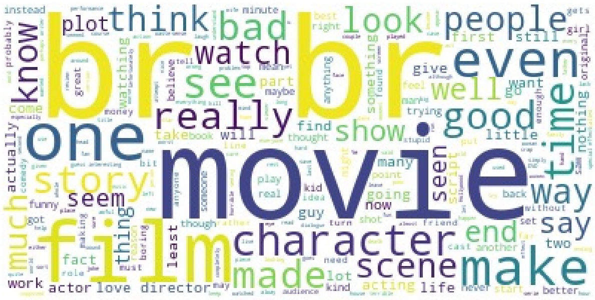


**Figure 6**: Word Cloud of Reviews Positive (left) and Negative (right)

Train and test sets are split based on the labels to train an LDA model. To visually see the frequency of words appearing in the reviews, two-word clouds are printed and shown in Figure 6. Since we want finer topics, words with low and high frequencies will be balanced through a TF-IDF matrix; non-negative matrix factorization (NMF), together with LDA, is used to extract topics. Here NMF is used to reduce dimensionality and solve the data sparsity problem. The top ten topic distribution of both positive and negative reviews are shown in Figures 7 and 8. Some similarities between the two, like 'movie(s),' 'film(s)' and 'story' frequently appear because all reviews, whether positive or negative, revolve around these contents and cannot contribute to sentiment analysis. As for positive reviews, words such as 'good,' 'like,' and 'great' frequently appear, which contain a more positive emotion. On the contrary, 'bad,' 'worst,' 'awful,' etc. appear in negative reviews, mostly complaining about some 'characters,' the 'director,' or the 'plot.' However, there are also positive topics appearing in negative reviews, meaning when people are reviewing the good or bad of a movie, they do not solely criticize it, but will judge it from multiple aspects to give a relatively objective point of view, and that is what makes this classification difficult.

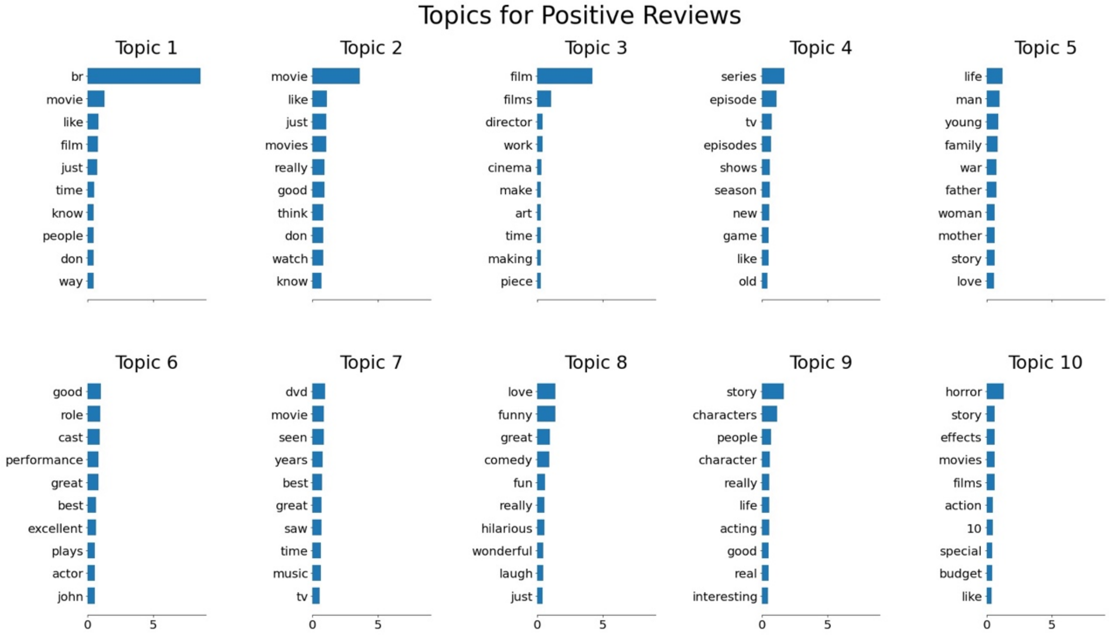


**Figure 7:** Topics for Positive Reviews

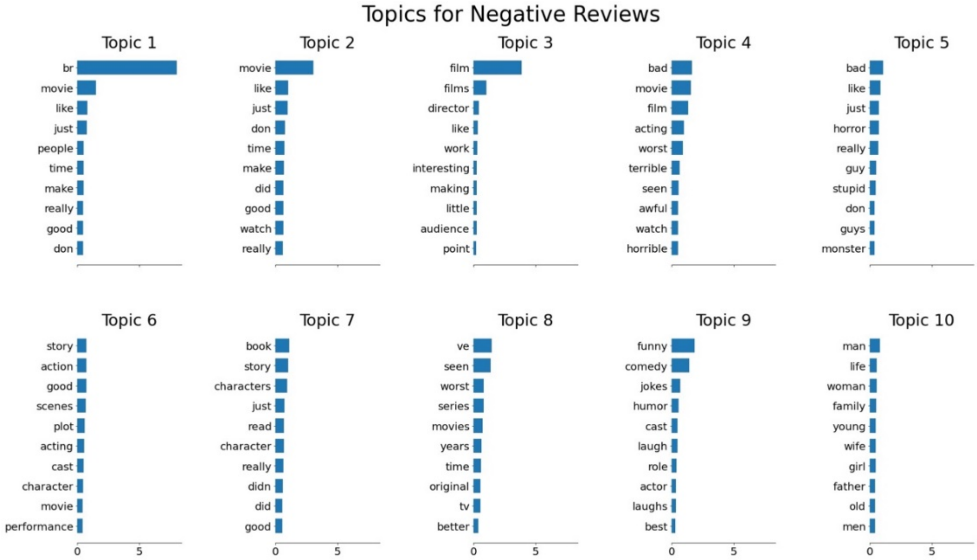


**Figure 8:** Topics for Negative Reviews

## 7. CONCLUSION

In this research, a BERT-CNN-based approach is applied to do sentiment analysis on movie reviews. Although its performance does not surpass the pure BERT model, it is proven that the BERT-CNN in this research is comparatively better at identifying negative sentiment, and it got an accuracy of around 94.4% after the ensemble. The topics of reviews with different views have also been investigated using LDA, showing that a movie review is usually multi-angle, so bringing local features (CNN) as support evidence could further confuse the model.

As a future direction, the max length of BERT can be increased to keep more details, and the CNN used should be more carefully designed to make pointed references to capture essential features. Also, since this research only focuses on text features, user characteristics like age, gender, or reputation are ignored. These factors also affect a person's view of things and should be considered.

## REFERENCES

1. Joyce, B. and Deng, J. (2017). Sentiment analysis of tweets for the 2016 us presidential election. In 2017 IEEE mit undergraduate research technology conference (urtc), pages 1–4. IEEE.
2. Drus, Z. and Khalid, H. (2019). Sentiment analysis in social media and its application: Systematic literature review. Procedia Computer Science, 161:707–714.
3. Ramos, J. et al. (2003). Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning, volume 242, pages 29–48. New Jersey, USA.
4. Das, B. and Chakraborty, S. (2018). An improved text sentiment classification model using tf-idf and next word negation. arXiv preprint arXiv:1806.06407.
5. Rhanoui, M., Mikram, M., Yousfi, S., and Barzali, S. (2019). A cnn-bilstm model for document-level sentiment analysis. Machine Learning and Knowledge Extraction, 1(3):832–847.
6. Sousa, M. G., Sakiyama, K., de Souza Rodrigues, L., Moraes, P. H., Fernandes, E. R., and Matsubara, E. T. (2019). Bert for stock market sentiment analysis. In 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), pages 1597–1601. IEEE.
7. Dong, J., He, F., Guo, Y., and Zhang, H. (2020). A commodity review sentiment analysis based on bert-cnn model. In 2020 5th International Conference on Computer and Communication Systems (ICCCS), pages 143–147. IEEE.
8. Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.
9. Jurafsky, D. and Martin, J. H. (2009). Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Pearson Prentice Hall, Upper Saddle River, N.J.
10. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
11. Zhang, R., Wei, Z., Shi, Y., and Chen, Y. (2019). Bert-al: Bert for arbitrarily long document understanding.