Analyzing film and drama reviews: Distinguishing trolls from genuine audience feedback based on the BERT model

Jiabei He

Information School, North China University of Technology, Beijing, 100144, China

21101020115@mail.ncut.edu.cn

Abstract. With the expanding influence of the Internet, an increasing number of individuals rely on viewer reviews to make informed decisions about whether to watch a movie or TV series. However, the prevalence of manipulated or "navy" reviews, employed by companies to boost their products' reputation, has created a significant challenge. While numerous studies have dissected film and drama reviews, a notable gap exists in discerning genuine audience feedback from deceptive ones. This article's research focus is on evaluating the model's capacity to effectively differentiate between authentic audience comments and navy reviews and delving into the complexities encountered when the model assesses comments, as well as highlighting the disparities between model-generated judgments and human assessments. This article first collects a large amount of different types of comment data, annotates these data, and then uses these data to train and fine tune the BERT model. Finally, the results are obtained and analyzed to determine the reasons. This article found that the accuracy rate of the model's judgment comments is around 71.08%, which is more accurate and stable. However, there are still some issues when judging comments with emojis and emoticons, and certain data is needed to support the judgment of comments for different movies or dramas. There are also certain issues with the dataset, as the data is manually annotated, and the annotation of the dataset itself may also be influenced by the annotator, which may lead to inaccurate judgments.

Keywords: BERT, film review, drama review, navy comments, audience comments.

1. Introduction

With the increasing development of the internet, an increasing number of people have the opportunity to watch various movies and TV shows online. As a result, there are also an increasing number of comments on various movies and TV shows on the internet, which can make it more convenient for people to determine the quality of a movie or TV show. However, many companies abuse the use of paid commenters to generate a large number of positive reviews for their movies or TV shows, in order to improve the ratings and attract viewers through false advertising. Initially, it was easy to distinguish between paid comments and genuine audience comments. However, a concerning trend has emerged where numerous companies exploit the use of compensated commentators to generate an abundance of favorable reviews for their productions. Their intention is to bolster ratings and lure in viewers through deceptive advertising practices.

This article believes that there is a need not only for a method to effectively distinguish comments, but also to provide an effective solution for the analysis of large-scale text data. In addition, due to the large number of reviews on movies and television programs, our work also has important practical value. Therefore, the purpose of this study is to use a data parallel computing framework and a BERT based analysis model to analyze and distinguish paid reviews and real audience reviews of movie and television program reviews, in order to achieve more efficient analysis. This article also aims to discover the differences between model judgment and human judgment in order to find ways to improve the model.

In recent years, some researchers have evaluated various aspects of movie reviews, such as some papers judging the revenue of movies based on the content of reviews [1]. However, more papers focus on the sentiment aspect of reviews, comparing the performance of supervised and unsupervised learning methods in sentiment analysis of movie reviews [2]. Some propose an aspect-based sentiment analysis model based on deep memory networks, which can explicitly model the connection between aspects and sentiments in reviews [3]. There are also papers that use pretrained Bidirectional Encoder Representation from Transformers (BERT) [4] models to extract aspect information from comments and perform sentiment classification for aspect sentiment analysis tasks [5].

Although there have been some studies analyzing movie and TV show reviews, these studies mainly focus on the content and information related to the reviews, or what the reviews express and analyze about the movies. They neglect the fact that reviews themselves may not contain useful information. In addition, most of these studies did not consider the specific requirements of movie and TV show reviews, such as the size of the corpus and the diversity of sentiment.

This research motivation comes from the challenges faced in analyzing movie and TV show reviews: how to efficiently distinguish between paid comments and genuine audience comments on a large-scale dataset. To address this issue, this article designs a BERT [4] model based on the Transformers [6] framework that uses data parallel computing. The method optimizes the BERT [4] model and then uses the data parallel computing framework for model training and judgment, while also identifying the differences between model judgments and human judgments on comments. Experimental results show that the proposed method can achieve high accuracy analysis on large-scale movie and TV show review data, and it significantly outperforms traditional single GPU methods.

2. Method

2.1. Dataset preparation

The dataset used in this article is extracted from comments from multiple movies and TV dramas by Douban. To ensure that the dataset covers a wide range of data types, this study selected 8, 000 comments from multiple types of movies and TV dramas such as fantasy, martial arts, plot, and romance as training and testing sets. When preprocessing the data, multiple annotators are invited to make judgments on the comments. Data considered to be audience comments is labeled with 1 in front of the text, while data considered to be navy comments is labeled with 0 in front of the text.

When annotating comments, certain data that poses a challenge in distinguishing between audience and navy comments will be circulated to a broader audience through questionnaires. The classification as either audience comments or navy comments will be determined through a voting process. Many difficult to judge comments have the following characteristics: comments are too brief (sometimes only a few words), comments have less information, comments have less emotion etc.

2.2. BERT

Bidirectional Encoder Representation from Transformers (BERT) is a state-of-the-art Natural Language Processing (NLP) model launched by Google in 2018 [4]. BERT has achieved significant results in various NLP tasks, including Q&A, emotion analysis, and language generation, thus revolutionizing the field of NLP. Figure 1 is the illustration of BERT model.

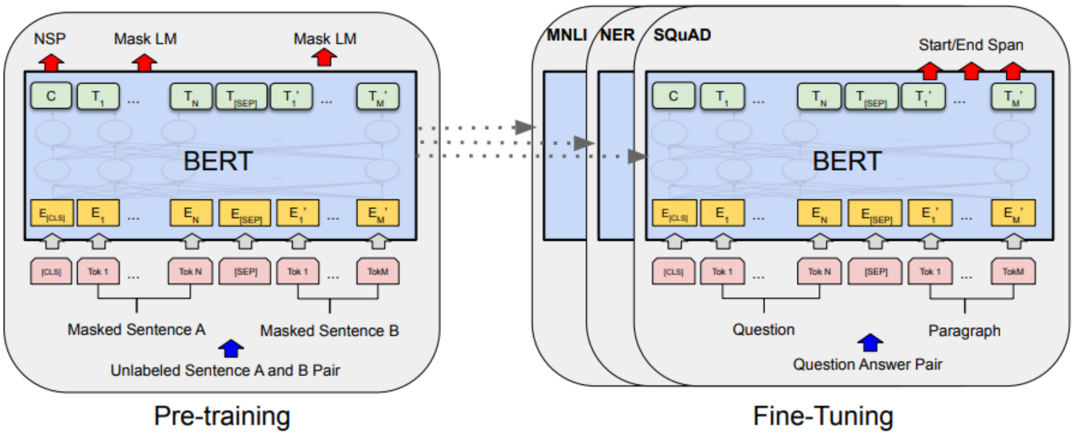

Figure 1. the illustration of BERT model [4].

The basic idea behind BERT is to use bidirectional encoding to capture contextual information of words in sentences. Unlike previous models that used unidirectional encoding, BERT processes input sentences from both left and right directions, enabling it to capture the complete context of each word. This method significantly improves BERT's ability to understand the meaning of words in sentences and generate more accurate predictions.

The basic principle of BERT is the self-attention mechanism, which enables the model to measure the importance of different words in sentences based on their contextual relationships. The attention mechanism allows BERT to dynamically adjust its attention to different parts of the input sentence, enabling it to capture the most relevant information for each word.

BERT consists of two main components: the embedding layer and the transformation layer. The embedding layer maps each word in the input sentence to a high-dimensional vector representation, capturing its semantic and syntactic characteristics. The conversion layer consists of multiple attention mechanisms and feedforward neural networks, enabling the model to learn complex relationships between words and sentences.

BERT's framework includes pre training models on large amounts of unlabeled text data using Masking Language Modeling (MLM) and Next Sentence Prediction (NSP) tasks. During the pre training process, BERT learns to predict missing words in sentences based on the surrounding context and generate coherent sentences with a given starting point. This process helps the model to gain a deeper understanding of language patterns and structures.

After pretraining, BERT fine-tuning can be performed on specific NLP tasks by adding task specific output layers. Fine tuning involves training models on labeled data for specific tasks, such as sentiment analysis or question answering, while maintaining pre trained parameters frozen. This enables BERT to utilize its pre trained knowledge while adapting to new tasks and domains.

2.3. Implementation details

This article makes some fine-tuning to the BERT [4] model, adjusting the learning rate to 1×10^{-5}, batchsize to 12, epochs to 10, and the optimizer uses AdamW [7, 8] to run the data in data parallel to accelerate the model.

3. Experimental results and analysis

From the Figure 2, it can be observed that after training, the accuracy of the model has been able to stabilize at 71.08%, and the accuracy in one round of training has basically increased steadily.

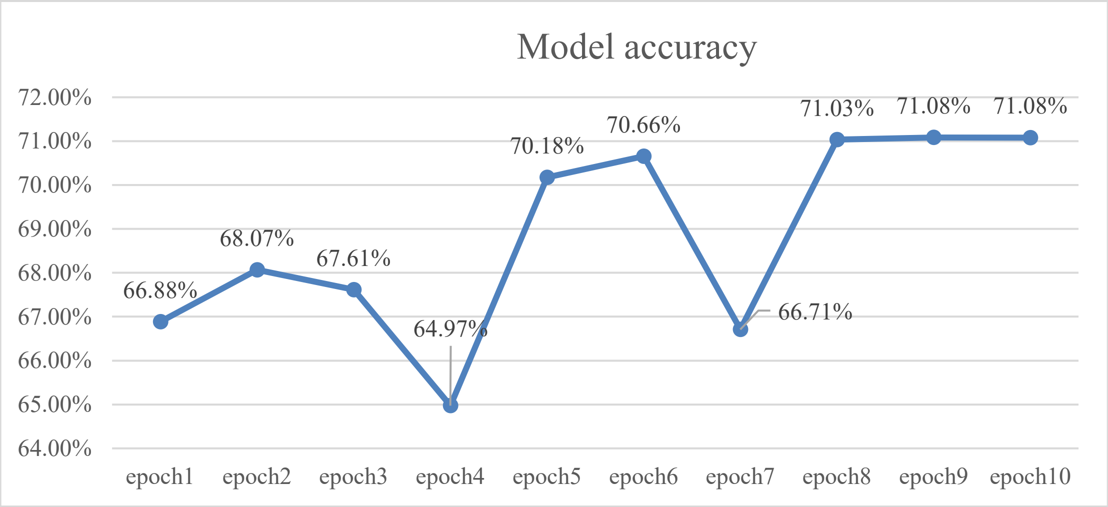

Figure 2. The change in accuracy of the model after each training session (Photo/Picture credit: Original).

However, the model still has a lot of room for improvement. Exploring the reasons behind the model's low accuracy and contrasting it with the disparity between humans and the model in assessing the authenticity of a movie review is the objective. This paper introduces softmax values [9, 10] to observe the accuracy of each comment judged by the model. When the softmax value [9, 10] is greater than 0.5, the model judges correctly; when it is less than 0.5, the judgment is incorrect. If the softmax value [9, 10] is between 0.4-0.5, it considers that the model's judgment on this comment is uncertain and belongs to ambiguous judgment.

3.1. Emojis and Emoticons

There are many emojis and emoticons in comments, and the error rate of the model when judging comments containing emojis and emoticons is 18%, which is lower than the error rate of the model when judging normal comments. Therefore, this paper deleted all the emojis in the comments before making a judgment, and the result in Figure 3 was 69.70%, which was very close to the model's judgment rate of 71%. When observing the softmax value [9, 10] of the model's judgment of the comments, it was found that only a few comments were ambiguous.

Furthermore, in cases where the model made an incorrect judgment on a comment containing an emoji or emoticon, removing the emoji or emoticon from the comment still carried an 83.92% probability of being assessed incorrectly.

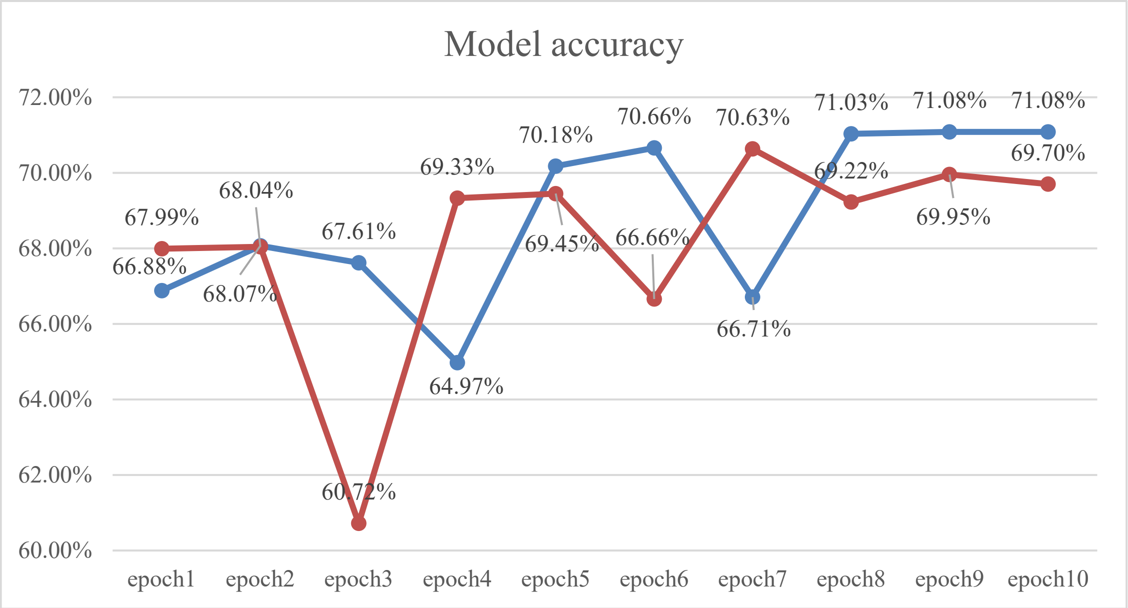

Figure 3. The blue line represents the original dataset, while the red line represents the dataset for removing emoticons and emoticons from comments (Photo/Picture credit : Original).

This article believes that this situation occurs because the model is unable to determine the specific meaning of the expressions and facial expressions in the comments. When annotating comments, annotators with emoticons and text in the comments tend to lean towards the audience's comments, as navy comments cannot reasonably use emoticons and text to express emotions. Only the audience can express emotions through emoticons and text. However, the model cannot distinguish the meaning of facial expressions and facial expressions, and even if the symbols of facial expressions and facial expressions are received, they can still only judge the text, only treating the facial expressions and facial expressions as one symbol. Therefore, the accuracy of the model in judging emoticon comments is similar to that of text comments.

This article also believes that another reason is that even if the model cannot distinguish the meaning of emoticons and emoticons, these symbols can still be considered as a characteristic of comments, and the feature of emoticons and emoticons in comments can be applied to different dramas or comments. Moreover, due to the inherent meaning of emoticons and facial expressions, when a comment contains these emoticons and facial expressions, the text itself may not contain more emotions or information, which leads to the model being unable to effectively judge these words after deleting the emoticons and facial expressions in the comment. By solely concentrating on comments featuring emoticons or facial expressions, the accuracy rate might not reveal any problems. However, when evaluating the overall accuracy rate, it tends to decrease.

3.2. Samples of Reviews for Each Movie or Series

During the training process, it is always believed that there is a big gap between paid comments and genuine comments, so if a comment model for a movie or series can distinguish whether it is a paid comments or not, the model can correctly judge other comments for another movie or series. This is feasible for annotators but it turns out that this is not correct for the model. As shown in Figure 4, the results of two trainings are compared, and it can be clearly seen that ACC2 has a higher and more stable accuracy rate, while ACC1 has a large fluctuation in accuracy rate and a lower accuracy rate.

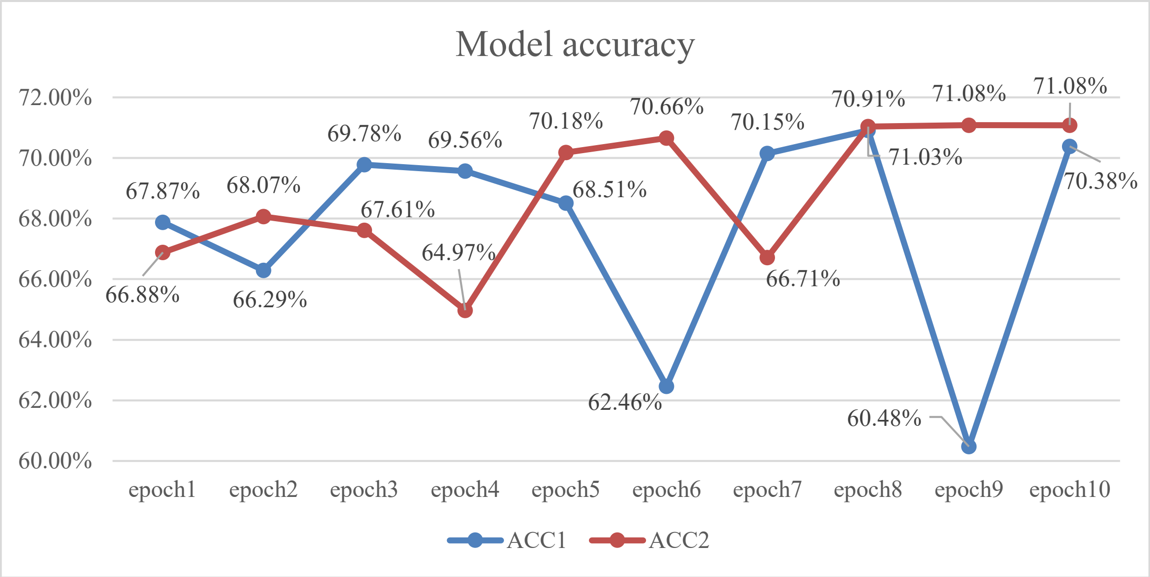

Figure 4. The red line represents the raw data, while the blue line represents the missing data in some navy comments (Photo/Picture credit : Original).

The reason for the difference in accuracy between the two training sessions is that the data from the first training session only included naval evaluations for movies and TV dramas, but lacked audience evaluations. The data from the second training session ensured that each movie and TV drama had two types of evaluations. Therefore, this article believes that in the model, it is necessary to determine whether the evaluation of a TV drama or movie is a water army, and the audience evaluation of the drama or movie should correspond to it, otherwise it will affect the accuracy of the model.

This article speculates that this is because each movie or TV series has proprietary words that are specific to that movie or TV series, and these words often appear in comments, such as the names of characters or actors in the play. These proprietary words do not appear in other TV dramas or movies, so the model cannot train to judge these words. If there are no comments from both parties for training, the model cannot judge comments with these newly appearing words, leading to judgment errors. The annotator possesses the ability to discern the meaning of these words, even in cases where they are unfamiliar with the noun's reference, such as a person's name or a place name. They can deduce this information from other contextual cues within the comment. In the comments, there may also be content that is completely unrelated to the drama or movie content. For the annotator, if these contents are completely unrelated to the movie or TV series, there is a high probability that it is a naval review, but the model cannot recognize these terms. Due to these issues, the content trained by the model in a movie or TV series review is not applicable in other movie or TV series reviews, which can also lead to a decrease in the accuracy of the model.

3.3. Discussion for Dataset Annotation

During dataset annotation, it has been found that batch labeling can lead to errors in judgment. The amount of information in a comment or the intensity of emotion is relative to other comments. Some comments may appear information-rich or emotionally intense when compared to others, but may not necessarily be perceived as such when judged independently. This situation has occurred multiple times during dataset annotation and cannot be completely avoided even with cross-judgment by multiple individuals or repeated evaluations of the dataset. Moreover, the judgment of comments is subjective, which means that there can be inherent issues in the annotation of the dataset itself.

4. Conclusion

Through experimentation, the study uncovered that the fine-tuned BERT model demonstrates effective discrimination between audience reviews and navy reviews in the context of film and drama assessments, consistently delivering high and stable accuracy in its evaluations. Nevertheless, the model faces challenges when appraising comments featuring emoticons and facial expressions. It struggles to interpret the intended significance of these emotive symbols, and the emotional nuances associated with emoticons and facial expressions may result in misjudgments, particularly when comments lack overt emotional expression. The model also needs some data to support the judgment of comments on different movies or dramas. If a movie or drama lacks comments from the navy or the audience, the model's judgment rate on the drama will decrease and the accuracy will be unstable. There are also certain issues with datasets, as the data is manually annotated, and annotators often annotate a large amount of data at once. Excessive comparison of data may lead to inaccurate judgment of comments by annotators. The main contribution of this article is to propose a method to distinguish whether a review is a naval review. On the one hand, it is beneficial for the audience to judge the true evaluation of a movie or drama, and on the other hand, it is beneficial for other researchers to exclude invalid data when analyzing film or drama reviews. There are still shortcomings in the current research on the judgment of comment symbols. This article only explores the symbols of emoticons and facial expressions, and sometimes multiple punctuation marks or other symbols may appear in comments. In the future, further refinement can be made on the relevant variables mentioned above to facilitate research on this topic.

References

[1] Mahesh J et al 2010 Movie Reviews and Revenues: An Experiment in Text Regression Proceedings of NAACL-HLT 2010 Short Papers Track.

[2] Bo P et al 2002 Thumbs up? Sentiment Classification using Machine Learning Techniques presented at the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP'2002)

[3] Duyu T et al 2016 Aspect Level Sentiment Classification with Deep Memory Network published in EMNLP 2016

[4] Devlin J et al 2018 BERT: Pretraining of deep bidirectional transformers for language understanding arXiv:1810.04805

[5] Manju V et al 2022 An enhanced guided LDA model augmented with BERT based semantic strength for aspect term extraction in sentiment analysis

[6] Vaswani A et al 2017 Attention is all you need 2017 in Proc Adv Neural Inf Process Syst pp 5998–6008

[7] Ye H et al 2018 Decoupled Weight Decay Regularization, Proceedings of the International Conference on Learning Representations (ICLR) arXiv:1711.05101

[8] Zhuang Z Liu M Cutkosky A and Orabona F 2022 Understanding adamw through proximal methods and scale-freeness arXiv preprint arXiv:2202.00089

[9] Bridle J S 1990 Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition Neural Computation 2(3) 227-236

[10] Hinton G E and Salakhutdinov R R 2006 Reducing the dimensionality of data with neural networks Science 313(5786) 504-507