Effectiveness of Multi-task Training for Prediction of Helpfulness of Online Movie Reviews

Che WANG Takuya MATSUZAKI
Department of Applied Mathematics, Faculty of Science Division I
Tokyo University of Science
1420019@ed.tus.ac.jp matuzaki@rs.tus.ac.jp

Abstract

This study examined the effectiveness of a multi-task learning for classifying review helpfulness, sentiment polarity, and ratio of sentences expressing positive sentiment in online movie reviews from IMDb. Employing a BERT-based model, the research demonstrates that integrating these tasks enhances model accuracy more effectively than handling them independently.

1 Introduction

The concept of helpfulness of online reviews, particularly that of movie reviews, is pivotal in guiding potential customers' purchasing decisions. In the digital era, consumers increasingly rely on reviews to assess the quality of products, especially on experiential products like movies. A helpful review typically contains detailed opinions and experiences, which are crucial for future sales and customer satisfaction. However, the enormous volume consisting of a mixture of good and bad reviews creates a challenge in discerning helpful from unhelpful content. Helpfulness is often quantified by the ratio of votes for helpfulness to total votes a review receives or by the total number of votes for helpfulness it garners. This metric is vital in e-commerce platforms, as it guides consumers in navigating through the vast amount of User-Generated Content (UGC) to find relevant and valuable insights about movies.

The prediction of review helpfulness is essential due to the overwhelming quantity of reviews on online platforms, which makes it difficult for potential customers to find helpful reviews. Automated helpfulness prediction can address the limitations of current systems, where new reviews may not have accumulated enough votes to reflect their true helpfulness. Predicting the helpfulness of reviews, particularly for new and unvoted content, can prevent the monopoly of high-ranked reviews and ensure that newer, potentially helpful reviews get noticed.

This study applies a multi-task learning (MTL) approach to the classification of review helpfulness, a method not extensively explored in this area previously. The incorporation of MTL, which enhances model performance and generalization by learning multiple related tasks concurrently, showcases the potential of this technique in comprehending and processing complex textual data. By concurrently tackling related tasks such as sentiment analysis and helpfulness classification, the model demonstrates improved capabilities in understanding the intricacies of textual data.

We consider both the overall sentiment polarity and the sentence-level positive ratio in reviews. While most existing studies focus solely on the overall sentiment polarity, this approach delves deeper by also accounting for the proportion of positive sentences in the total number of sentences in reviews.

The remainder of the paper is structured as follows. Section 2 summarized related work. Section 3 details the method, followed by the experimental setup. This is succeeded by the results and discussion, and the paper concludes with the Conclusion section.

2 Related Work

In recent years, some scholars have conducted research on the prediction of the helpfulness of reviews and explored influencing factors. Hao et al. [1] explored the factors affecting the helpfulness of online reviews from the perspective of text features, finding that positive emotional tendencies and higher levels of mixed positive and negative emotions in online movie reviews have a significant positive impact on the perceived helpfulness. However, their approach involved treating these factors as variables in a regression analysis of review helpfulness, and the small data size limits the generalizability of the results. Liu et al. [2] used a similar method to demonstrate that positive extremity in reviews positively affects their helpfulness. Han et al. [3] simply explored the linear relationship between ratings and review helpfulness, showing a weak positive correlation, but their study employed statistical methods rather than NLP techniques. Xu et al. [4] proposed a BERT model that incorporates ratings as feature vectors to predict review helpfulness, demonstrating its effectiveness. Bilal et al. [5] discovered that fine-tuned BERT-based classifiers outperform bag-of-words approaches in classifying helpful and unhelpful reviews. Alsmadi et al. [6] presented a set of deep learning models to predict the helpfulness of online reviews, but did not consider factors influencing review helpfulness. Chua et al. [7] found a certain relationship between ratings and review helpfulness, relying on statistical procedures of analysis of variance and multiple regression, but the authors do not claim causality. Singh et al. [8] developed machine learning models that proved average rating, entropy, and sentiment parameters to be important factors in review helpfulness. Saumya et al. [9] used a CNN model to preserve semantic features of sentences, proving the effectiveness of a 2-CNN model in predictions.

3 Method

This section consists of two parts: task descriptions and multi-task learning framework.

3.1 Task Descriptions

Task 1: Helpfulness Ratio Three-Class Classification
The first task is to predict the helpfulness ratio of a review, which is quantified as the number of votes for helpfulness divided by total votes. The primary objective is to classify the helpfulness ratio into three distinct categories, assessing the model's performance through the accuracy of this three-class helpfulness classification.

Task 2: Sentiment Polarity Binary Classification
For the second task, the model will be trained using review texts and their associated overall sentiment polarities. These polarities are derived from the ratings provided in the reviews.

Task 3: Positive Ratio Three-Class Classification
The third task involves predicting the ratio of positive sentences within reviews. Initially, an ALBERT model [10], pre-trained on the SST-2 dataset with 93% accuracy, predicts the sentiment polarity of each sentence. We then calculate the ratio of positive sentences to the total number of sentences in a review. The multi-task model will then be trained using review texts and their associated positive ratios to accomplish the three-class positive ratio classification.

3.2 Multi-task Learning Framework

Initially, as a baseline to evaluate the effectiveness of multi-task learning, a BERT-based model is used to independently perform the first task, which is the three-class classification of the helpfulness ratio. This step serves as an initial reference point for comparison, aiming to clearly demonstrate the performance enhancements brought about by multi-task learning.

In the two-task setting, the study implements a combination of the second and third tasks with the first task separately to observe the enhancement in model performance. In this setup, the model shares its initial layers up to the output of the CLS vector for both tasks. This means that the input and hidden layers of the model, which capture the common features and semantic information, are shared across tasks. Subsequently, the path diverges into two feedforward networks, where each task is processed separately with different parameters. This design aims to assess the potential for performance improvement on specific tasks by sharing underlying representations.

In the three-task setting, the model integrates all three tasks for simultaneous training. Similar to the two-task setup, the model separates the processing of each task in the feedforward network part. By doing so, the model concurrently considers the classification of the helpfulness ratio, the overall sentiment polarity, and the distribution of the sentence-level polarity, aiming at maximizing the synergistic effect among these tasks. The purpose of this integrated multi-task learning approach is to explore the impact of different task combinations on the overall performance of the model, particularly in terms of accuracy and generalization capability.

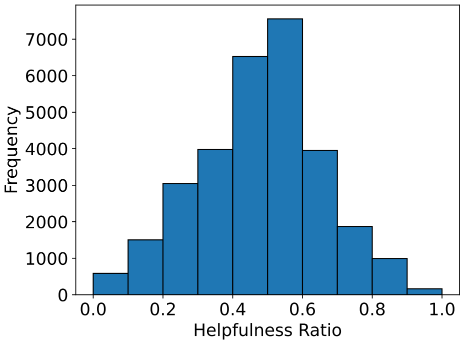

Figure 1 Distribution of Helpfulness Ratio

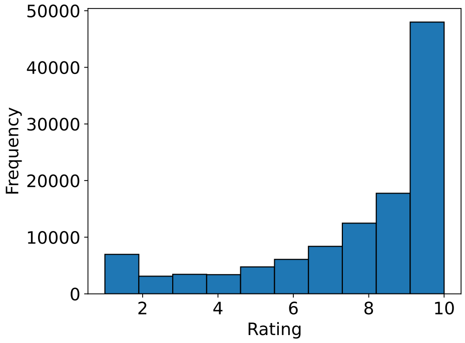

Figure 2 Distribution of Rating

4 Experimental Setup

4.1 Data Preparation

The training and the test data was collected from the IMDb website,[1] specifically from the "Most Popular Movies" section as of November 22, 2023. The dataset includes the top 50 movies sorted by the number of ratings. For each movie, the dataset comprises the review text, vote data (including votes for helpfulness and total votes), user ratings for the movies, and the overall movie ratings.

4.2 Experiment Setup

Initially, a series of data preprocessing steps were conducted on the dataset obtained from the IMDb website, which included the cleansing of text data. Subsequently, entries without user ratings were first eliminated. Given that a small total number of votes in the calculation of the helpfulness ratio can lead to extreme outcomes that affect training effectiveness, only data with total votes equal to or greater than 10 were used in the first task's dataset (amounting to a total of 30,175 entries). The reviews were then categorized into three classes based on the distribution of the helpfulness ratio (as shown in Figure 1): 0–0.4, 0.4–0.6, and 0.6–1. The ratio of the class sizes was 33%, 44%, and 23%, respectively.

For the second and third tasks, the datasets were not restricted by the total number of votes (totaling 114,372 entries). In the second task, based on the overall distribution of ratings (as shown in Figure 2), the reviews with ratings equal to or greater than 9.0 were classified as 1 (positive), with all others categorized as 0 (negative).The ratio of the class sizes was 43% and 57%, respectively.

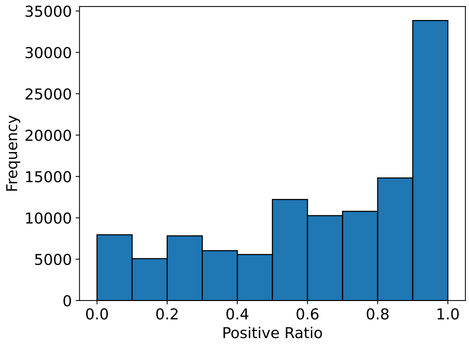

Figure 3 Distribution of Positive Ratio

In the third task, the sentiment polarity of each sentence was initially predicted using the ALBERT model as explained in 3.1. This was followed by calculating the ratio of positive sentences to the total number of sentences in each review. According to the distribution of positive ratio (as shown in Figure 3), the data was segmented into three classes: 0-0.5, 0.5-0.8, and 0.8-1. The ratio of the class sizes was 33%, 32%, and 35%, respectively.

The experiment initially conducted Task 1 separately on the BERT model to assess its accuracy, serving as a baseline for evaluating the effectiveness of subsequent multi-task learning. Similarly, Tasks 2 and 3 were individually conducted to observe their standalone performances. Following this, a two-task and a three-task learning approach were initially implemented without setting specific loss weights. This was done to observe the model's performance in a more straightforward multi-task learning context. Subsequently, loss weights proportional to the size of the respective datasets were established for a re-evaluation (1:3.79 for two-task learning and 1:3.79:3.79 for three-task learning). This weighted approach aimed to balance the influence of each task based on dataset size, assuming that a

[1] https://www.imdb.com/chart/moviemeter/?ref_=nv_mv_mpm

Table 1 Accuracy of Task(s)

	Task(s)	Helpfulness	Polarity Sentiment	Ratio Positive
	1	0.4983		
	2		0.8579	
	3			0.7580
No Loss Weight	1+2	0.5514	0.8626	
	1+3	0.5524		0.7557
	1+2+3	0.5550	0.8514	0.7485
Set Loss Weights	1+2	0.5319	0.8655	
	1+3	0.5506		0.7561
	1+2+3	0.5471	0.8636	0.7500

larger dataset may require more attention during the training process. In both phases of the experiment, the model shared parameters before the feedforward network, ensuring that the initial layers, capturing the common features and semantic information, were utilized across all tasks. Finally, the outcomes of the multi-task learning were compared with the results obtained from conducting the tasks independently.

For determining the number of learning steps, 10% of the training set data was reserved as a development set. Evaluation utilized parameters from the epoch with the highest accuracy on the development set within each task. Maximum number of epochs was set to 10.

5 Main Results and Discussion

Independent Task Outcomes

Table 1 presents the evaluation results of the single-task, two-task and three-task settings. The initial task of classifying the helpfulness ratio of reviews into three categories yielded an accuracy of 49.83% on the test set. This task served as a baseline for assessing the impact of multi-task learning.

For Task 2, the model achieved an accuracy of 85.79% in sentiment polarity classification. Task 3 showed a performance of 75.80% in positive probability classification.

Enhancements through Multi-Task Learning

The transition to multi-task learning scenarios unveiled significant findings. When Tasks 1 and 2 were combined without setting loss weights, the helpfulness accuracy improved markedly to 55.14%. Similarly, the combination of Tasks 1 and 3 led to an accuracy of 55.24% for helpfulness. Notably, integrating all three tasks (1+2+3) further improved the helpfulness accuracy to 55.50%, while maintaining strong performance in sentiment polarity (85.14%) and positive ratio (74.85%). This increase suggests that learning shared representations across tasks can provide deeper insights, particularly for complex tasks like helpfulness classification. The combined learning approach seemed to enable the model to leverage the nuances of overall sentiment polarity and sentence-level positive ratio in understanding helpfulness, a synergy not tapped in isolated task learning.

Reflections on Loss Weight Adjustments

The introduction of loss weights, calibrated to the size of the respective datasets, did not yield a consistent improvement in performance. In cases of the 1+2 and 1+2+3 task combinations, the performance with loss weights (accuracy of 53.19% and 54.71% respectively) was slightly less favorable compared to the same combinations without loss weights (accuracy of 55.14% and 55.50% respectively). This suggests that while the concept of loss weight adjustment is theoretically sound, its practical application might require more nuanced considerations, particularly in balancing the contributions of each task in a multi-task learning setting.

6 Conclusion

This research examined the effectiveness of multi-task learning for classifying review data. The results demonstrated the effectiveness of this approach, highlighting the benefits of leveraging shared representations across related tasks. However, the results were not uniformly strong across all multi-task settings. Additionally, the introduction of loss weights, intended to proportionally balance the influence of each task based on dataset size, did not consistently enhance performance.

This outcome highlights the complexities and challenges inherent in predicting user perceptions based solely on textual data. The modest improvements observed suggest that while multi-task learning can be a valuable tool in natural language processing, there is significant room for refinement and optimization in model architecture and training methodologies.

References

[1] 郝媛媛, 叶强, 李一 军. 基于影 评数据的在 线评论有用性影 响因素研究. 管理科学学报, Vol. 13, No. 8, pp. 78–88, 2010.

[2] Zhiming Liu, Li Hong, and Lu Liu. An investigation of on-

line review helpfulness based on movie reviews. **African Journal of Business Management**, Vol. 8, No. 12, pp. 441–450, 2014.

[3] 韩雁雁, 张宁, 房文敏. 基于 meta 分析的在 线评论有用性影 响因素模型研究. 信息系 统学报, No. 1, pp. 1–13, 2015.

[4] Shuzhe Xu, Salvador E Barbosa, and Don Hong. Bert feature based model for predicting the helpfulness scores of online customers reviews. In **Advances in Information and Communication: Proceedings of the 2020 Future of Information and Communication Conference (FICC), Volume 2**, pp. 270–281. Springer, 2020.

[5] Muhammad Bilal and Abdulwahab Ali Almazroi. Effectiveness of fine-tuned bert model in classification of helpful and unhelpful online customer reviews. **Electronic Commerce Research**, Vol. 23, No. 4, pp. 2737–2757, 2023.

[6] Abdalraheem Alsmadi, Shadi AlZu'bi, Mahmoud Al-Ayyoub, and Yaser Jararweh. Predicting helpfulness of online reviews. **arXiv preprint arXiv:2008.10129**, 2020.

[7] Alton YK Chua and Snehasish Banerjee. Helpfulness of user-generated reviews as a function of review sentiment, product type and information quality. **Computers in Human Behavior**, Vol. 54, pp. 547–554, 2016.

[8] Jyoti Prakash Singh, Seda Irani, Nripendra P Rana, Yogesh K Dwivedi, Sunil Saumya, and Pradeep Kumar Roy. Predicting the“helpfulness”of online consumer reviews. **Journal of Business Research**, Vol. 70, pp. 346–355, 2017.

[9] Sunil Saumya, Jyoti Prakash Singh, and Yogesh K Dwivedi. Predicting the helpfulness score of online reviews using convolutional neural network. **Soft Computing**, Vol. 24, No. 15, pp. 10989–11005, 2020.

[10] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. **arXiv preprint arXiv:1909.11942**, 2019.