Comparative Sentiment Analysis of Movie Reviews Generated by Critics and General Audience

Zijin Nie
Faculty of Engineering
McGill University
260787728

Hao Shu
Faculty of Engineering
McGill University
260776361

Chenzhun Huang
Faculty of Science
McGill University
260761859

{zjin.nie,hao.shu,chenzhun.huang}@mail.mcgill.ca

https://github.com/VincentCloud/COMP550-final-project

Abstract

Sentiment analysis aims to predict and determine the opinion of a writer from textual data. By applying this technique to reviews, one can gain precious insights into people's attitudes towards a certain item. As one of the typical sentiment analysis datasets, movie reviews were most commonly classified as a whole in previous work, while little research investigated reviews generated by different types of people. In this research, we collected a dataset of 14K movie reviews separated into critic-generated and audience-generated categories from RottenTomatoes. We observed that the general audiences' reviews gave better classification accuracy and conducted experiments on this phenomenon. We showed that the distribution of opinion words and review length had a great effect on sentiment analysis and resulted in the discrepancy of prediction accuracy between the two types of reviews.

1 INTRODUCTION

With a massive amount of data generated on the internet, automatic text classification has drawn great interest. Movie reviews from the public provide insights into movie productions. Sentiment analysis could come in handy for such a use case. Websites for movie reviews such as Rotten Tomatoes have gained immense popularity.

With the advancement in deep learning and modern neural networks, the performance of sentiment-analysis models is taken to the next level. DL-based models also became robust enough to tackle more-advanced variations like fine-grained sentiment analysis and intent analysis. Linear regression is the most straightforward but less satisfactory solution. And numerous models have been proposed to achieve high performance on these classification tasks. One such example would be BERT (Bidirectional Encoder Representations from Transformers).

Movie critics tend to focus more on evaluating the techniques and artistry rather than expressing subjective opinions straightforwardly. The formality and style of reviews could have an effect on the sentiment analysis performance. In preliminary tests, we found that both the Logistic Regression and BERT(Devlin et al., 2019) model performed better in reviews from the general audience than ones from critics.

We proposed three factors that could result in this discrepancy: the use of opinion words, the review length, and the sentence composition i.e POS tags. Experiments showed audience reviews on average used two times more opinion words than critics reviews. By removing the opinion lexicon from our dataset, the difference margin dropped from 5% to 2%. In addition, we showed that longer reviews were easier to predict by comparing reviews above and below length median on critic and audience data respectively. However, the proportion of POS tags in the two categories did exhibit noticeable relation with sentiment prediction.

2 RELATED WORK

Sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. Sentiment analysis systems have been applied to many kinds of literatures including novels (Boucouvalas, 2003), emails (Liu et al., 2003), news articles (Samuels and Mcgonical, 2020), tweets in Twitter (Pak and Paroubek, 2010), and movie reviews (Baid et al., 2017) (Sahu and Ahuja, 2016). Previous works have explored sentiment analysis on different types of literature. Kaur et al (Kaur and Saini, 2014) show that the level of text formality could significantly impact the performance of sentiment analysis. However, we found very little research investigated the formality and style difference of the

same type of text, particularly when the literature is generated by amateurs and professionals.

In opinion mining of reviews, sentiment words and part-of-speech tagging are regarded as very common techniques. For instance, Hu et al. (Hu and Liu, 2004) automatically selected a list of most important features from the training corpus and achieved a strong prediction precision against these words. Text length might be an important factor in classification performance. In 2010, Wang et al. performed experiments on the effect of text length on the classification performance (Wang and Dong, 2010). The results showed that longer sentences may improve the classification performance due to the more information provided. However, if the sentences were too long, the performance could be negatively affected since the sentence might introduce contradicting information.

3 BACKGROUND

3.1 BERT

Builds on top of the transformer architecture, BERT is the state-of-the-art contextual pre-training method. Models built on top of BERT achieved superior performance in a wide range of NLP tasks including sentiment analysis, question answering and machine translation. As a general-purpose language representation model, BERT can be fine-tuned on a specific dataset or task to gain accuracy performance using minimum effort. (Miller, 2019).

3.2 Sentiment Score

In 2014, Kiritchenko et al. proposed the method to compute the sentiment score for a term *w* in short informal textual messages (Kiritchenko et al., 2014):

$$Sentiment\ \ Score\,(w) = PMI\,(w, positive) - PMI\,(w, negative)$$

where *PMI* stands for pointwise mutual information (Kiritchenko et al., 2014):

$$PMI\,(w, positive) = log_2\,\frac{freq\,(w, positive) * N}{freq\,(w) * freq\,(positive)}$$

where *freq (w, positive)* stands for the number of times a term w occurs in positive messages, *freq (w)* stands for the total frequency of term *w* in the corpus, *freq (positive)* stands for the total number of tokens in positive messages, and N is the total number of tokens in the corpus.Therefore, the sentiment score could be computed as:

$$Sentiment\ \ Score\,(w) = log_2\,\frac{freq\,(w, positive) * freq\,(negative)}{freq\,(w, negative) * freq\,(positive)}$$

A positive sentiment score indicates a greater overall association with positive sentiment. In contrast, a negative score indicates a greater association with negative sentiment.

4 EXPERIMENT APPROACH

To set up the experiment, we collected movie reviews from both the audience and the critics. Then we obtained preliminary results using Logistic Regression and BERT trained on two sources. Based on the results, we proposed three hypotheses that could potentially lead to the performance gap between the two datasets. To evaluate the hypotheses, we conducted various experiments focusing on the effect of sentiment words, text lengths and POS composition.

4.1 Data Collection

There are numerous existing movie review datasets taken from RottenTomatoes, but we could not find any data that separates the reviews from different sources. Therefore, we decided to extract such information from the website directly.

Using the python web crawling framework, we scraped 13439 movie reviews from the RottenTomatoes website. The critic group had 5457 positive reviews and 3449 negative reviews, respectively. The audience group had 3187 positive reviews and 1346 negative reviews, respectively. Critic reviews on Rotten Tomatoes mostly reference external websites, and each review shows only a portion of a whole article to convey the critics' point of view on the movies to a large extent. The complete reviews from the audience are stored directly inside the website and are shown in full. Therefore, the reviews from the audience were directly collected.

The rating scales from the two categories differ slightly. For critics' reviews, there were two ratings, one is taken from the website and the other uses the Rotten Tomatoes' standard "fresh" or "rotten". The reviews from the audience range from one star to five with a scale of 0.5. Since our study only involves binary sentiment analysis, we decided on a custom metric to align the two reviews and label them in a binary manner. For the critics, existing measures were converted into positive ("fresh") or negative ("rotten"). While for the audience, we

ruled all the ratings including and below 3 stars to be negative reviews because for most of the reviews the positive polarity only shows above 3 stars.

4.2 Experiment Setup

For preliminary testing, we ran a Multiclass Logistic Regression model and a pre-trained English uncased bert_L-4_H-512_A-8 model (Devlin et al., 2019) to evaluate the sentiment analysis performance on our dataset. Both models are implemented with the scikit-learn and Keras library in TensorFlow. The BERT model is fine-tuned on the movies review data to achieve a better performance than the default.

We proposed several metrics to quantify the performance difference. To expose the internal characteristics of data, we analyzed the average length and the POS composition of both types of reviews.

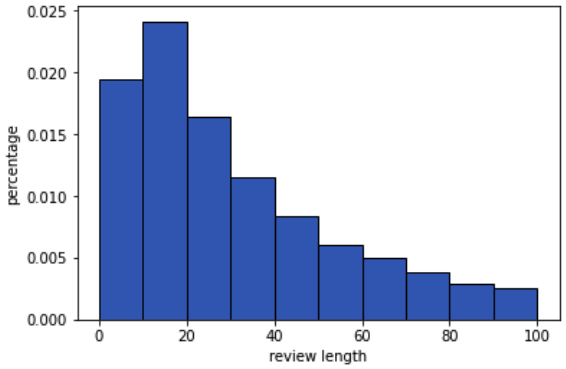
4.2.1 Hypothesis I: Effect of Opinion Words

Our first hypothesis is that the general audience tends to use more opinion words than film critics, which makes the movie review from one group easier to predict than the other. Mohammad (2012) showed that emotion words such as "surprised", "disgusted" are good indicators that the text as a whole is expressing the same emotion. To keep professionalism and conscientiousness, film critics focus more on evaluating the techniques and artistry of movies. Their careful wording could make the opinions more obscure.

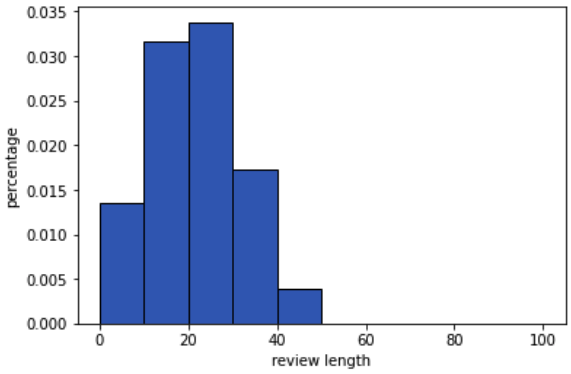
To test the theory, we first filtered out a list of tokens with a sentiment score(Kiritchenko et al., 2014) lower than -5 or higher than 10. We selected these thresholds manually to achieve a reasonable list size and to cover the words that are commonly used as sentiment cues. These tokens exhibit a high correlation to one of the binary labels. Excluding stop-words, a set of total 335 tokens were chosen. Then we investigated the distribution of sentiment words among both types of reviews by calculating the total occurrences of these tokens separately.

However, the token distribution alone does not strongly justify the effect of these words. On top of that, we decided to mask the strongest sentiment words in both types of reviews and see whether it compensates for the difference in sentiment analysis accuracy. We modified the original dataset by replacing all occurrences of the opinion lexicon with the "_UNKNOWN_" token. Then we obtained the performance on the new dataset by retraining the Logistic Regression and BERT models.

4.2.2 Hypothesis II: Review Lengths



(a) Length distribution for audience reviews



(b) Length distribution for critic reviews

Figure 1: Length Distribution for Reviews from Audience and Critics

Figure 1 presents the distribution of the review lengths from critics and audience, respectively. We can observe that the length of the critics' reviews ranges from under 10 words to 50 words, with [10, 20] and [20, 30] being the majority of the review lengths. However, for the audience reviews, there is a much larger presence in the intervals of longer review length, even beyond the specified range. The average length of the audience review text is a striking 61 compared to (??) for the critics' reviews. And the median lengths for each category are 29 and 22, respectively. We have reasons to believe that the length of the review text might also be a factor contributing to the performance of the sentiment analysis model in the two groups. More specifically, in an idealized situation, a longer polarized review could exhibit more traits in favor of a certain sentiment, which could improve the classification confidence of our model.

To test this hypothesis, we separated each group according to their median review lengths. Then the logistic regression model was trained on the data

with longer (group 1) and shorter reviews (group 2), respectively. The model was then trained on half of all the reviews regardless of length. As a comparison, we then randomly sampled half of the data points from reviews with all lengths (group 3).

4.2.3 Hypothesis III: Parts of Speech

Our third hypothesis is that the percentages of part-of-speech tags in sentences would also affect the sentiment classification results. During the experiment of the prior hypothesis, we noticed that the majority of the sentiment words that had high sentiments were adjectives and verbs. Previous work by Wang et al. (Wang et al., 2018) and Kalarani et al.(Kalarani and Brunda, 2019) used part-of-speech tagging as an important feature for sentiment analysis. The experiments performed by Karamibekr et al. (Karamibekr and Ghorbani, 2012) also showed that verbs could be a better determination of sentiment to improve classification results.

To test this hypothesis, we tagged every category of part-of-speech tag in the critic review and audience review, by using the nltk part-of-speech tagging tool (Bird et al., 2009). Then the percentage of each part-of-speech was computed based on the total number of words in the sentences.

5 RESULTS

5.1 Effect of Removing Sentiment-related words

We calculated the total occurrences of sentiment words in audience and critic review separately. Table x shows that every critic review contains 5.07 opinion words on average, while every audience review uses 15.44 opinions words averagely.

Table 2 shows the prediction accuracy before and after removing the set of words with high sentiment scores from datasets. The average prediction accuracy of the baseline is 84.5% on the audience review and 79.8% on the critic review. There was a 5% difference margin between the two in the original data. After removing the sentiment words(noted as x-no-sent in Table 2), the margin value dropped to 2% (74.7%/76.8%) on Logistic Regression(LR) and BERT models, which was reduced by around 60% compared to the original result. It was obvious that the absence of these sentiment words penalized the prediction accuracy of the audience's reviews much more than the critics'.

	# of tokens	# of reviews	Average
Critics	9367	47520	5.07
Audience	4533	69990	15.44

Table 1: Average token count of opinion words used in each sentenced

	LR	BERT
Audience	84.34%	85.11%
Critics	79.27%	80.43%
Audience-no-sent	76.96%	76.68%
Critics-no-sent	74.79%	74.67%

Table 2: Classification accuracy of critics and audience movie reviews before and after removing sentiment words

	LR	BERT
critics_long	80.80%	77.10%
critics_short	74.51%	76.99%
critics_half	76.63%	81.37%
audience_long	83.63%	83.33%
audience_short	79.74%	78.83%
audience_half	82.82%	75.31%

Table 3: Classification accuracy on reviews with lengths above and below median

5.2 Review Lengths

As seen from Table 3, the accuracy of both models trained on reviews above the median length is generally higher than those trained on reviews below the median length. And models trained on randomly sampled reviews generally have the accuracy between the previous two.

5.3 Experiment on Parts of Speech

After tagging the part-of-speech in critic reviews and audience reviews. The computed percentage of each category is shown in Table 4. Unexpectedly, there was no significant difference in the POS makeup of the two types of reviews.

6 DISCUSSION AND CONCLUSION

Our experiments about the sentiment words showed that the general audience tends to use more emotional words in movie reviews than professional film critics. Given an unequal distribution of polarized words, the sentiment analysis models were more accurate on the set of text that has more sentiment cues. We are convinced that the effect of sentiment cues are universal across models because

POS	Critics	Audience
Verb	15.828%	17.294%
Noun	29.837%	28.238%
Adjective	11.277%	9.984%
Adverb	6.695%	6.828%
Others	36.363%	37.656%

Table 4: POS composition of audience and critic review in percentage

the phenomena were similar on both the statistical Logistic Regression model and neural-based BERT model,

The test on Hypothesis II suggests that our sentiment analysis models tend to have better accuracy on movie reviews with a larger word count. However, the setup of hypothesis II is not comprehensive. With a longer review text, there could also be more tokens present that are associated with the opposite sentiment. With that limitation in mind, we should refine our hypothesis that, with a longer text, there is more increase in the occurrences of tokens with the same polarity than those of the opposite.

In the investigation of the POS features, we found that the reviews from the audience and critics had similar POS tag compositions. It means that sentiment has hardly any relation with the sentence structure and grammatical tagging of the corpus of our proposed dataset. However, we did observe that adjectives and verbs appear more frequently in the list of opinion words with high sentiment scores.

6.1 Conclusion and Takeaway

In this paper, we explored the factors that could contribute to the discrepancy in the performance of binary sentiment classification. We proposed three hypotheses on the difference in the use of opinion words, the length of the reviews, and part-of-speech compositions of reviews between the two groups. We found a higher usage of relatively strong opinion words from the audience group, and a similar performance of models trained on the dataset with those words removed, which supported the first hypothesis. In addition, both of the models showed superior performance on longer review texts than the shorter ones, which implies the correlation between text length and sentiment classification accuracy. Although justified by various previous works studying the relations between POS and sentiment analysis model, there is no significant difference between the POS composition of the reviews from the two groups.

6.2 Further development

The current dataset has an imbalanced positive/negative data ratio for both critic and audience review. To make our claim more convincing, we need to expand the dataset by adding more negative reviews to eliminate the data bias. With our limited computational resources, our dataset only includes partial critic reviews. A more powerful web crawler could be applied to extract full critic reviews and provide more training examples. With more data entries and longer critics reviews, the dataset can be applied to other types of topics and it will be more suitable for neural-based approaches which usually require a huge amount of data to converge.

In addition, we can expand the research using other datasets. For example, we can compare the restaurant reviews on Yelp generated by both general dinners and professional food critics using almost the same approach. These further explorations could potentially strengthen our hypotheses and provide more insights.

7 STATEMENT OF CONTRIBUTIONS

All members have made equal contributions towards this project. The distribution of work for each member is described as follows:

Zijin Nie : Implementing BERT model and extracting sentiment words

Hao Shu : Data preprocessing, logistic regression classification, test on part-of-speech hypothesis.

Chenzhun Huang : Data collection and test on the review length hypothesis

References

Palak Baid, Apoorva Gupta, and Neelam Chaplot. 2017. Sentiment analysis of movie reviews using machine learning techniques. *International Journal of Computer Applications*, 179:45–49.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Anthony Boucouvalas. 2003. 21 real time text-to-emotion engine for expressive internet communica-

tions. *International Journal of Communication Systems - Int. J. Communication Systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *Proceedings of the 19th National Conference on Artifical Intelligence*, AAAI'04, page 755–760. AAAI Press.

P. Kalarani and S. Selva Brunda. 2019. Sentiment analysis by pos and joint sentiment topic features using svm and ann. *Soft Computing*, 23:7067–7079.

Mostafa Karamibekr and Ali A. Ghorbani. 2012. Verb oriented sentiment classification. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 327–331.

Jasleen Kaur and Jatinderkumar Saini. 2014. Emotion detection and sentiment analysis in text corpus: A differential study with informal and formal writing styles. *International Journal of Computer Application*, 101:1–9.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *J. Artif. Int. Res.*, 50(1):723–762.

Hugo Liu, Henry Lieberman, and Ted Selker. 2003. A model of textual affect sensing using real-world knowledge. IUI '03, page 125–132, New York, NY, USA. Association for Computing Machinery.

Derek Miller. 2019. Leveraging bert for extractive text summarization on lectures.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. volume 10.

Tirath Prasad Sahu and Sanjeev Ahuja. 2016. Sentiment analysis of movie reviews: A study on feature selection amp; classification algorithms. In *2016 International Conference on Microelectronics, Computing and Communications (MicroCom)*, pages 1–6.

Antony Samuels and John Mcgonical. 2020. News sentiment analysis.

Jianxiong Wang and Andy Dong. 2010. A comparison of two text representations for sentiment analysis. *ICCASM 2010 - 2010 International Conference on Computer Application and System Modeling, Proceedings*, 11.

Yili Wang, Kyung Tae Kim, Byung Jun Lee, and Hee Yong Youn. 2018. Word clustering based on pos feature for efficient twitter sentiment analysis. *Human-centric Computing and Information Sciences*, 8:1–25.