Dropping Fake Favorable Feedback for Better Sentiment Analysis

Qiankun Su, Zhida Feng, Wujing Sun, and Lizhen Chen†
College of Computer Engineering, Jimei University, Xiamen, China
{qiankun.su, 202011810014, 201721121010, lzchen}@jmu.edu.cn

ABSTRACT

Fake favorable feedback is a big obstacle for customers to get better experiences on shopping websites. As far as we know, the effects of fake favorable feedback on natural language processing models have not been reported. To investigate the distraction of fake feedback, this paper developed a method to resolute fake feedback (RFF). The proposed RFF first analyzes the tokenizes of feedback, and then several short texts are generated to replace some long texts. Experimental results on the raw data and processed data show the effectiveness of our proposal.

CCS CONCEPTS

• Computing methodologies • Machine learning • Machine learning approaches • Neural networks

KEYWORDS

Chinese Sentiment Analysis; BERT; Customer feedback; Fake Favorable Feedback

1 Introduction

E-commerce carries considerable weight in the global retail framework. Global e-retail sales reach 4.2 trillion U.S. dollars in 2020[1]. China has the largest digital buyer population in the world. Its e-commerce sales surpassed the combined total of Europe and the United States[2]. Customers are likely to share their experience with a product or service in the form of customer feedback, including the price and quality of goods, shopping experience. This feedback is of great importance for consumers, retailers, e-commerce platforms, and manufacturers.

Sentiment analysis is widely applied to customer feedback to identify, extract, quantify, and study affective states (such as negative, moderate, positive). Consumers, retailers, e-commerce platforms, and manufacturers use this valuable information to make better decisions. A significant percentage of consumers regularly check out ratings and feedback of the product they want to buy.

Various approaches for sentiment analysis have been developed[3-5]. Harnessing the power of deep learning, some research leverages deep learning to sentiment analysis and achieves much better results recently. There are two most well-known models, BERT (Bidirectional Encoder Representations from Transformers) [6] and GPT-3[7] (Generative Pre-trained Transformer 3). GPT-3 contains 175 billion tunable parameters. Running it consumes large amounts of memory and compute resources[8]. For this reason, we leverage BERT for sentiment analysis in this paper. More specifically, we applied the existing pre-trained BERT model that can be fine-tuned with one additional output layer to create a new model for our sentiment analysis.

A high-quality dataset plays an important role in training the model. Usually, customer feedback as the training dataset is collected from e-commerce websites. However, such a dataset existing plenty of fake feedback, especially favorable feedback. A product is more likely to show up in the search results if it gets a higher score with more positive feedback. Naturally, the product is more likely to be purchased. Merchants make more profits. Therefore, merchants have a strong motivation to encourage consumers to leave favorable customer feedback. Incentives include coupons, red packets, cashback. The feedback is usually provided by the merchants and then simply copied and left by the consumers. Even worse, some merchants employ the services of a click farm. These measures make positive feedback be flooded with fake favorable feedback.

To dispel the effects of fake favorable feedback, we firstly investigate the effects of padding size on the accuracy of our proposed model using the raw dataset. We notice that an obvious feature of fake favorable feedback is long texts. Based on this finding, we drop some fake favorable feedback and create some positive feedback with short texts to replace them. With the new processed dataset, we compare the performance of our model on the raw dataset and the processed data. The result shows that dropping favorable feedback indeed improves the accuracy of our model and meanwhile consumer less computing resources.

This paper is organized as follows. Section 2 summarizes the main related work. The proposed model is presented in Section 3. Experimental results and their analyses are given in Section 4. Finally, Section 5 concludes this work.

2 Related Work

Existing approaches to sentiment analysis can be classified into two basic categories from a technical point of view: lexicon-based techniques and machine learning approaches[3]. The lexicon-based methods run fast, but with relatively low accuracy. Machine learning approaches achieve higher accuracy but require a huge dataset to train the model. OpenAI recently published GPT-3[7]. It is regarded as the world's most sophisticated natural language technology at present. GPT-3 has 175 billion parameters. It is trained with 499 billion tokens. To avoid large-scale parameter learning for each task, a wide range of pre-trained models have been developed[9]. Once a model is pre-trained, it can be shared to save a lot of memory and computational power. One of the latest milestones in this development is the release of BERT[6] by Google. It is a transformer-based architecture that uses a self-

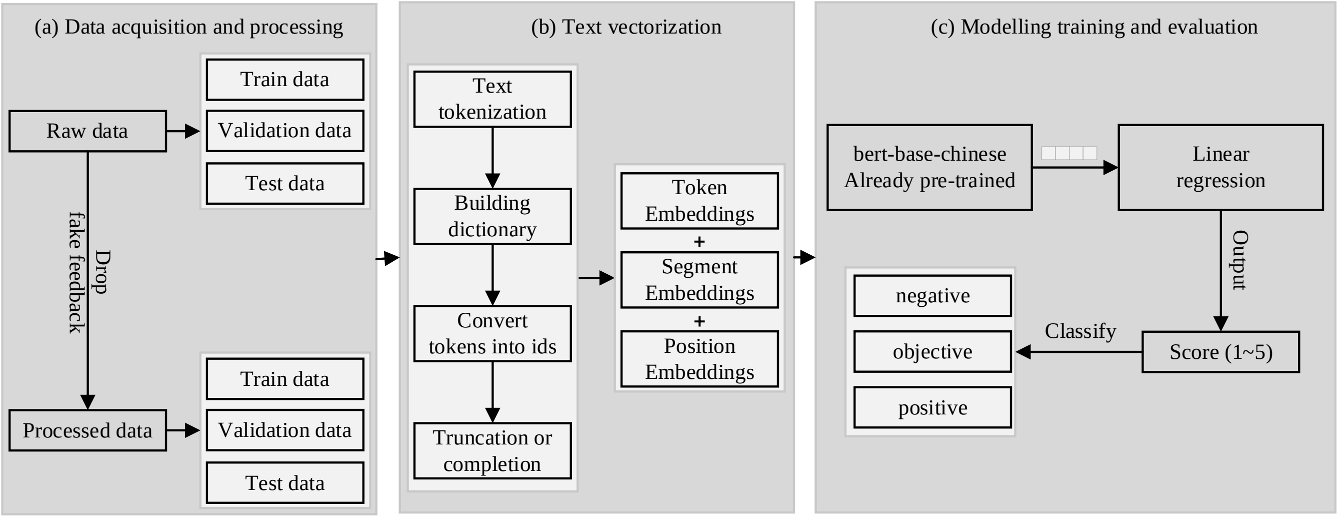


Figure 1: The graphical diagram of the proposed RFF model.

attention mechanism. BERT makes full use of both left and right contexts in all layers. It consists of two phases, pre-training followed by supervised task-specific fine-tuning. For this reason, we leverage the BERT pre-trained model for our task in this paper.

Most research on sentiment analysis focuses on English language. It is still not mature for Chinese language[10]. The biggest difference between them is no space between words in Chinese language. The granularity levels in Chinese sentiment analysis could even be character level. In [11], the authors conduct experiments to show that character-based models consistently outperform word-based models. This is because the sparse distribution of Chinese words is likely to result in overfitting. For this reason, we carry out character-level sentiment analysis for customer feedback.

Some research has been conducted on Chinese sentiment analysis on feedback[12-16]. Most of them are based on the BERT model. However, they did not consider fake feedback. In this paper, we investigate the effects of fake favorable feedback.

3 The Proposed RFF

This section illustrates the proposed model (RFF) to resolute fake feedback for Chinese sentiment analysis. The diagram of RFF is depicted in Figure 1. Firstly, we acquire customer feedback from a Chinese e-commerce website, hereafter referred to as "Raw data". To eliminate the effects of fake favorable feedback, we create a new dataset, hereafter referred to as "Processed data", in which some of the fake favorable feedback are replaced with some short favorable feedback generated automatically. Each dataset is divided into three parts, train, validation, and test data. Secondly, we process feedback text into vectors as the input for the BERT model. Thirdly, we leverage the existing pre-trained BERT model, BERT-Base-Chinese[1] for our task.

We add a linear regression layer to model the relationship between feedback and score. To further investigate the performance of our model, we categorize customer feedback into negative, moderate, and positive feedback. Finally, we evaluate the performance of our proposed method.

3.1 Dropping Fake Favorable Feedback

To make more profit, merchants have a strong motivation to encourage consumers to leave favorable feedback. Incentives include coupons, red packets, cashback. The feedback is usually provided by the merchants and then simply copied and left by the consumers. Even worse, some merchants employ the services of a click farm. Therefore, the positive feedback in the raw dataset contains an abundance of fake favorable feedback. It causes the model trained inaccurately. We notice that most of the real feedback is short texts. Therefore, we drop some fake favorable feedback and add the positive feedback with short texts to the dataset.

The way we create favorable feedback with short texts is as follows. Firstly, we summarize the features that consumers focus on by analyzing customer feedback. For instance, consumers evaluate milk products from the following aspects: taste, smell, price, manufacture date, package, and speed of delivery. Secondly, we have adjectives to describe the above features. Taking the feature 'taste' as an example, the adjectives would be good, nice, not bad, delicious, etc. Thirdly, we have adverbs to qualify the adjectives, such as very, greatly. Finally, with the summarized nouns, adjectives, and adverbs, we permutate over (nouns, adverbs, adjectives) and combinate the generated short sentences. We use the created favorable feedback to replace fake

[1] https://huggingface.co/bert-base-chinese

favorable feedback partly in raw data. Thus, we have a new dataset, referred to "Processed data" in the following sections.

3.2 Text Vectorization

We convert our data to tensors as an input format for BERT. Firstly, we break a feedback text into a list of tokens. Secondly, it can build a dictionary from all customer feedback and assign a vocabulary ID to each token. With the dictionary, we convert tokens into ids. Thirdly, to ensure the same length of input vectors, we truncate the list of vocabulary indices if the length of tokens is greater than the assigned padding size, or fill up with MASK vice versa. Finally, the input embedding is the sum of the token embeddings, the segment embeddings, and the position embeddings. The input representation is passed to BERT's attention layer.

3.3 Model Training

BERT model contains two key components, pre-trained BERT model and fine-tuning BERT model. There are many variants of pre-trained model. We use bert-base-chinese developed by Hugging Face. For the fine-tuning BERT model, we add a linear layer to model the relationship between feedback text and score. To further investigate the effects of fake favorable feedback on the accuracy of our model, we use a simple classifier to categorize feedback into negative, moderate, and positive feedback.

4 Results and Analyses

In this section, we evaluate the performance of our method in terms of MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), R^2 (Coefficient of determination), and accuracy. In short, we make the following observations:

1. In section 4.1, we find that the key feature of fake favorable feedback is having much more characters.
2. In section 4.3 and section 4.4, we confirm that existing fake feedback has a side effect on the model.
3. In section 4.4, we create a new dataset by dropping some fake favorable feedback. The model trained with the processed data outperforms the model trained with the raw dataset. The accuracy improves 21.2%, from 76.5% to 97.7%. Meanwhile, the model achieves the best performance much faster.

4.1 Analyses of Raw Dataset

We select customer feedback of dairy products to evaluate our model for the following reasons: i) Milk products are the food of mass consumption and therefore abundant feedback are available; ii) Commendatory and derogatory terms are obvious in customer feedback of daily products. The vast majority of them are blunt or straightforward, with very little irony feedback.

We obtain customer feedback from JD.com, one of the two massive B2C online retailers in China by transaction volume and revenue. The feedback is labeled as one star to five stars. We evenly scrape 29,000 positive, moderate, and negative feedback respectively, together with corresponding ratings. A 3-star rating is regarded as moderate feedback. More and less than represents positive and negative feedback respectively.

Figure 2 presents the histogram and cumulative distribution function (CDF) of negative, moderate, positive, and all customer feedback. The number of characters of the largest number of negative and moderate feedback is found at 12, however, the number is 105 for the positive feedback. More interestingly, 89.6% of negative feedback has no more than 22 characters. 91.4% of moderate feedback have no more than 22 characters. However, the number is only 60.3% for the positive feedback.

Obviously, positive feedback has much more characters than negative or moderate feedback in general. Most of them are fake favorable feedback.

4.2 Dataset and Experimental Setup

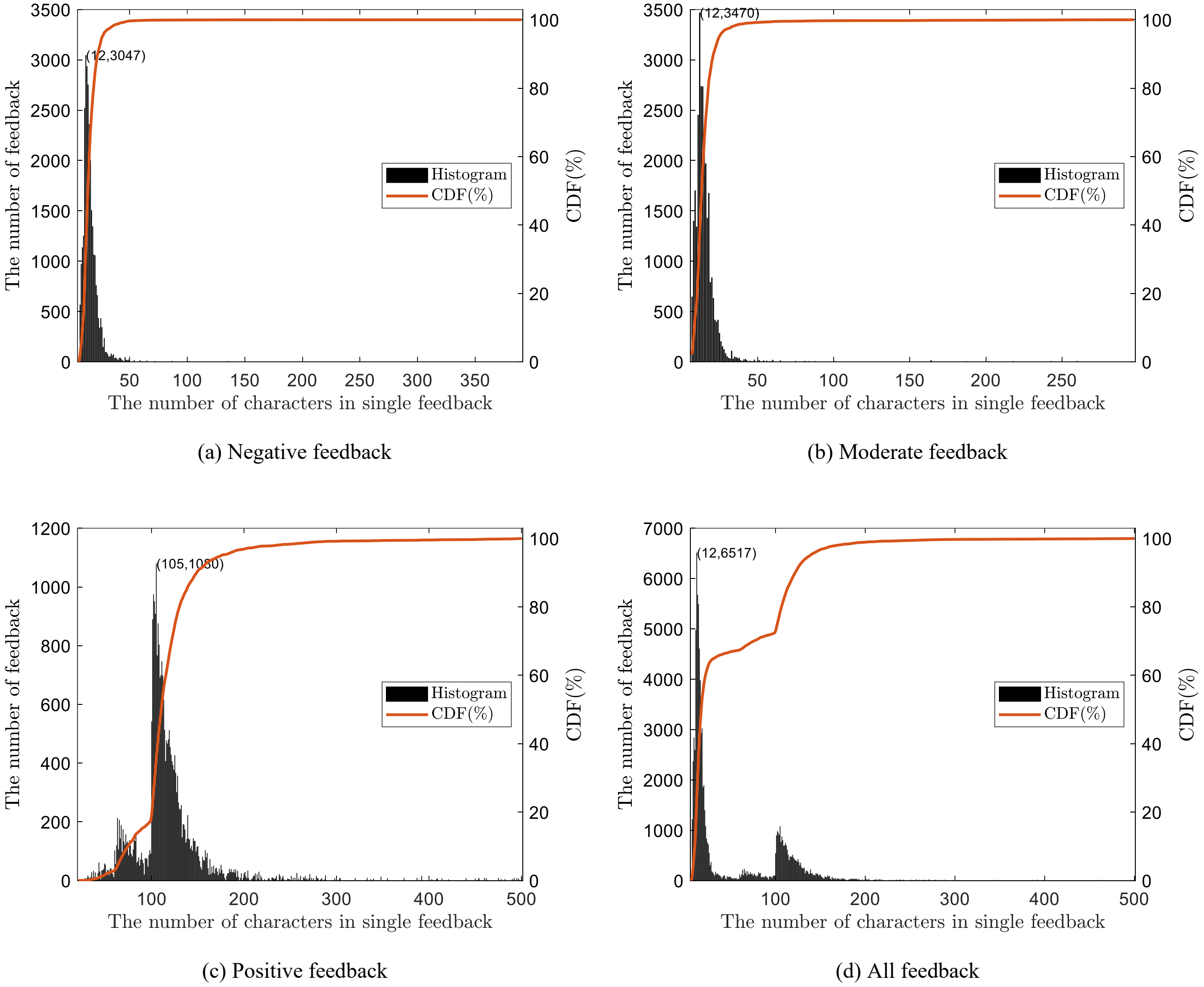
Raw dataset. The raw dataset is obtained as described in Section 4.1. The orderly arrangements of customer feedback lead to move in the direction of the gradient regularly. It might be difficult to find the optimal weights. To avoid this, we shuffle the feedback randomly. After this, we split the dataset into train, validation, and test data, the first 90% of data as train data, the following 5% as test data, and the remaining 5% as validation data. The test set is referred to as "Test1".

Test set 2. As explained in Section 4.1, the positive feedback in the raw dataset contains plenty of fake favorable feedback. To investigate the performance of our model RFF on real feedback, we scrape favorable feedback with short texts to replace that of Test1. The new test set is referred to as "Test2".

Processed dataset. To dispel the side effects of fake feedback, we generate an abundance of positive feedback with fewer characters to replace with raw positive feedback. Specifically, we generate 14084 feedbacks, almost half of the total positive feedback. Thus, the processed positive feedback contains both long and short texts. This new dataset is referred as to "Processed data". Similar to raw datasets, the processed dataset is divided into train, validation, and test data. To evaluate the performance of RFF on real feedback, we use Test2 as the testing set instead.

Performance metrics. We evaluate our proposal with the widely used metrics, MAE, RMSE, R^2, accuracy, and confusion matrix.

Configurations. We split the dataset into train, validation, and test data, the first 90% of data as train data, the following 5% as test data, and the remaining 5% as validation data. The learning rate is set to 1e-5, and the batch size is set to 64. The dropout rate is set to 0.1 for all the models.



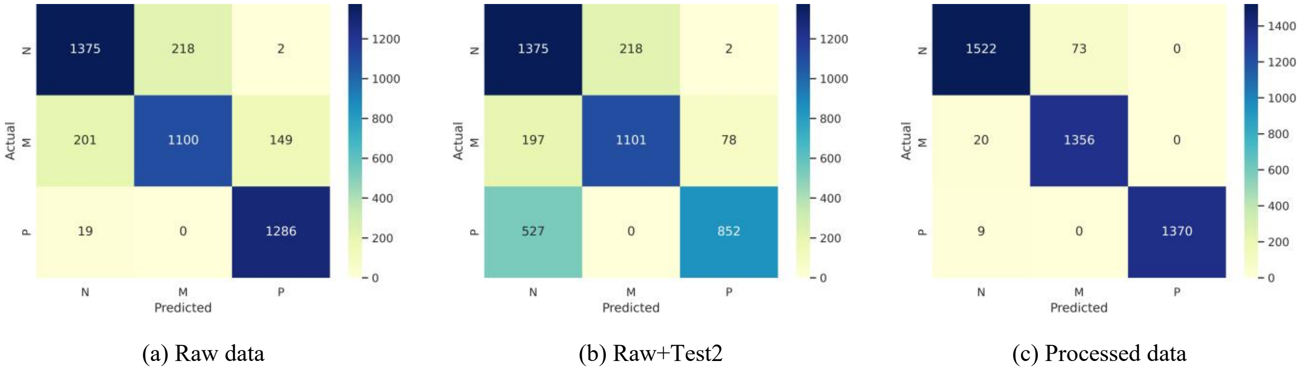
(a) Negative feedback

(b) Moderate feedback

(c) Positive feedback

(d) All feedback

Figure 2: The histogram and CDF of negative, moderate, positive, and all customer feedback.

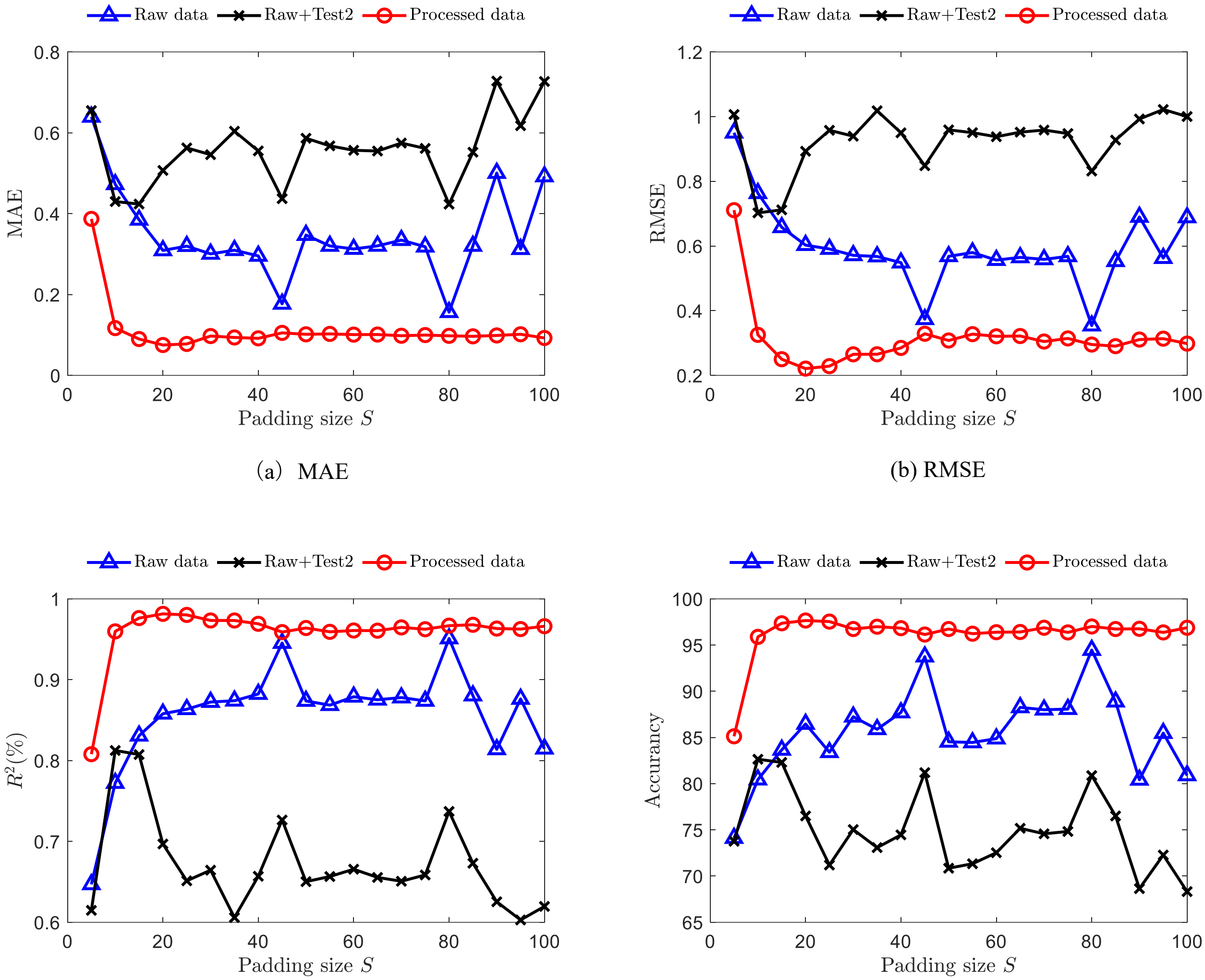


(a) Raw data

(b) Raw+Test2

(c) Processed data

Figure 3: The comparisons of confusion matrix among raw data, Raw+Test2, and processed data. (The padding size is fixed at 20.)



(a) MAE

(b) RMSE

(c) R^2

(d) Accuracy

Figure 4: The experimental comparisons among raw data, raw data testing on real feedback, and processed data in terms of MAE, RMSE, R^2, and accuracy.

4.3 Impact of the Padding Size

We evaluate the effects of padding size on the performance of our model in terms of MAE, RMSE, R^2 and accuracy. The blue line in Figure 3 illustrates MAE, RMSE, R^2 and accuracy over a series of padding sizes. In the beginning, with the increase of the padding size from 5 to 40, MAE and RMSE decrease steadily; R^2 and the model accuracy increases steadily. As the padding size continues to increase, these values remain stable. There are some anomalous points, such as the padding size 45. The potential reason is that the distribution of the number of characters of positive feedback is quite different from that of negative or moderate feedback.

4.4 The Effects of Fake Favorable Feedback

We use the model trained with raw data to evaluate scores of positive feedback with short texts. Most of their score is 3-stars and even lower, i.e., falling into the category of moderate feedback and even negative feedback. To confirm this, we use a new test set whose positive feedback is real ones, i.e., short texts, to test on the above trained model. The experimental results are plotted in Figure 3, labeled as "Raw+Test2" in black. It shows that the pattern is similar to that of testing on the raw test set. Not surprisingly, MAE and RMSE are much higher while R^2 and accuracy are much lower than that of testing on the raw test set. The highest accuracy is only 82.6%, which is found at the padding size 10. Figure 4 reveals this point in a more visually. It shows the confusion matrix of raw data, raw data testing on real feedback at the padding size 20. As we can see from Figure 4-(b), 527 positive feedback (on the lower left) improperly falls into negative feedback.

As we stated in Section 4.1, positive feedback is flooded with fake favorable feedback that has more characters. It causes that the trained model is in favor of the positive feedback with long texts.

4.5 Dispel the Effects of Fake Favorable Feedback

To dispel the side effects of fake feedback, we process the raw dataset by generating an abundance of positive feedback with fewer characters and then replacing some of the raw positive feedback.

Figure 4 presents the comparison of the performance of RFF on raw data, raw dataset testing on real favorable feedback, and processed data in terms of MAE, RMSE, R^2 and accuracy. Clearly, the model with the processed data achieves much better performance. Take the padding size 20 as an example, compared with Raw+Test2, the model trained with processed data (labeled as "Processed data" in red) achieves a decrease of 85.18% and 75.26% in MAE and RMSE respectively, an increase of 40.84% in R^2. The accuracy improves 21.2%, from 76.5% to 97.7%. When it comes to the confusion matrix as depicted in Figure 3-(c), there is very little positive feedback mistakenly. More interestingly, the model with processed data achieves the best performance with a much smaller padding size than that of Raw+Test2, which means it runs much faster and saves computing resources.

5 Conclusion

In this paper, we leverage BERT for sentiment analysis on customer feedback in Chinese. We find that large quantities of fake favorable feedback exist in customer feedback by evaluating the effects of the padding size on our method. Consequently, we propose a method to drop fake favorable feedback for better sentiment analysis. Extensive experiments show the validity of our proposal.

ACKNOWLEDGMENTS

This work is supported by Jimei University (no. ZQ2018008), the Education Department of Fujian Province (CN) (no. JAT200273), and the Natural Science Foundation of Fujian Province (CN) (no. 2019J05099). The authors would like to thank the editor and anonymous reviewers for their insightful feedback on improving the manuscript quality.

REFERENCES

[1] Statista. 2021. Worldwide e-commerce share of retail sales 2015-2024. (July 2021).

[2] Statista. 2021. E-commerce in China. (May 2021).

[3] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. Knowl. Based Syst. 226 (2021), 107134. https://doi.org/10.1016/j.knosys.2021.107134.

[4] Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 8, 4 (2018). https://doi.org/10.1002/widm.1253.

[5] Doaa Mohey El-Din Mohamed Hussein. 2018. A survey on sentiment analysis challenges. Journal of King Saud University - Engineering Sciences 30, 4 (2018), 330–338. https://doi.org/10.1016/j.jksues.2016.04.002.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423.

[7] OpenAI. GPT-3: Language Models are Few-Shot Learners. [EB/OL]. https://github.com/openai/gpt-3 Accessed June 13, 2021.

[8] Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its Nature, Scope, Limits, and Consequences. Minds Mach. 30, 4(2020), 681–694. https://doi.org/10.1007/s11023-020-09548-1.

[9] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-RongWen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. Pre-Trained Models: Past, Present and Future. CoRR abs/2106.07139 (2021). arXiv:2106.07139 https://arxiv.org/abs/2106.07139.

[10] Haiyun Peng, Erik Cambria, and Amir Hussain. 2017. A Review of Sentiment Analysis Research in Chinese Language. Cogn. Comput. 9, 4 (2017), 423–435. https://doi.org/10.1007/s12559-017-9470-8.

[11] Xiaoya Li, Yuxian Meng, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li. 2019. Is Word Segmentation Necessary for Deep Learning of Chinese Representations?. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 3242–3252. https://doi.org/10.18653/v1/p19-1314.

[12] Mingzheng Li, Lei Chen, Jing Zhao, and Qiang Li. 2021. Sentiment analysis of Chinese stock reviews based on BERT model. Appl. Intell. 51, 7 (2021), 5016–5024. https://doi.org/10.1007/s10489-020-02101-8.

[13] Yalin Miao, Wen Fang Cheng, Yi Chun Ji, Shun Zhang, and Yan Long Kong. 2021. Aspect-based sentiment analysis in Chinese based on mobile reviews for BiLSTM-CRF. J. Intell. Fuzzy Syst. 40, 5 (2021), 8697–8707. https://doi.org/10.3233/JIFS-192078.

[14] Song Xie, Jingjing Cao, Zhou Wu, Kai Liu, Xiaohui Tao, and Haoran Xie. 2020. Sentiment Analysis of Chinese E-commerce Reviews Based on BERT. In 18th IEEE International Conference on Industrial Informatics, INDIN 2020, Warwick, United Kingdom, July 20-23, 2020. IEEE, 713–718. https://doi.org/10.1109/INDIN45582.2020.9442190.

[15] Li Yang, Ying Li, Jin Wang, and R. Simon Sherratt. 2020. Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning. IEEE Access 8 (2020), 23522–23530. https://doi.org/10.1109/ACCESS.2020.2969854.

[16] Huibing Zhang, Junchao Dong, Liang Min, and Peng Bi. 2020. A BERT Fine-tuning Model for Targeted Sentiment Analysis of Chinese Online Course Reviews. Int. J. Artif. Intell. Tools 29, 7-8 (2020), 2040018:1–2040018:23. https://doi.org/10.1142/S0218213020400187.