Customer Satisfaction Research based on Customer Service Dialogue Corpus

Jian Chai¹, Shengfu Wang², Jie Zhu²*, Xian-Ling Mao² and Heyan Huang²

¹ China Mobile Online Service Co., Ltd., Henan, China
² Beijing Institute of Technology, Beijing, China

* Email: zhu_j@bit.edu.cn

Abstract. Recently, consumers are increasingly inclined to contact customer service for help when they encounter problems and have high demands for remote support. A high-quality customer service connects the company with its customers and establishes a positive image. Applying customer satisfaction metrics to measure the quality and efficiency of customer service is widely used, yet most of the existing customer service evaluation systems rely on manual processes, which are clearly unsustainable and costly. We introduce an ERNIE-based customer satisfaction analysis model that automatically analyses the text of customer service dialogues and scores them from four perspectives (i.e., product, service, process and overall) without human involvement. Furthermore, we construct a corpus containing around 1500 entries of dialogues texts transcribed from customer service consultation and scale it up to 9 times in the training phase. Results show that our model performs better compared to the baseline model and demonstrates a good generalization ability as well.

1. Introduction
In recent years, with the rapid development of China's economy and society, people's consumption ability is constantly increasing. Amidst the fast-paced life, more and more consumers are willing to contact customer service for assistance when they encounter problems with products. As a result, consumers have a high demand for remote service, which is especially critical in the current pandemic outbreak. A high-level customer service connects the company with its customers, builds a positive public impression, and ultimately achieves the long-term goal of fostering consumer loyalty to the company and its products [1][2]. Therefore, measuring customer satisfaction is becoming the core of every customer-centric company.

Customer satisfaction metrics provide the company with valuable insights into their overall business performance, yet there is no universally accepted way to measure it [3]. Most existing customer service evaluation systems rely on customers to score the customer service they have experienced in person, which is quite cumbersome for those desperate for help. Also, other systems rely on third-party evaluation agencies manually scoring the satisfaction metrics of each consultation, which is clearly unsustainable and costly.

Therefore, in efforts to reduce the reliance on manual intervention for evaluating customer satisfaction metrics, we treat the task of evaluating customer satisfaction as a text classification task and introduce an ERNIE-based customer satisfaction analysis model. The proposed model automatically analyses the text of a conversation between the customer and the customer service representative and scores it from four perspectives (i.e., product, service, process, and overall) without
human intervention. For training the text classification model based on ERNIE [4], we construct a corpus of around 1,500 (scaled to 9 times the size in the training phase) transcribed text of customer service consultation dialogues, and finally obtain the customer satisfaction from four different perspectives mentioned above. Experimental results indicate that our proposed model achieves better F1 score in all four perspectives compared to two baseline models, by at least 1 to 2 percentage points.

The contributions of this work can be summarized as follows:

- We construct a corpus of approximately 1,500 transcribed texts from real-world customer service consultation dialogues, containing labels of customer satisfaction in four perspectives: product, service, process, and overall. Each dialogue text is annotated as satisfied or unsatisfied according to uniform annotation principles. We do our best to ensure that the annotation results truly reflect customer service satisfaction by considering contextual information.
- We use the ERNIE-based text classification model to evaluate customer satisfaction, which takes a pre-trained ERNIE model as its core. As a sentence-level feature, the representation of the [CLS] token is fed into a softmax classifier to yield the posterior for satisfaction and unsatisfaction.
- Experimental results show that our approach is way better than other two baseline models. After analysis, we conclude that service satisfaction is the most vital part of customer satisfaction, and product and process also have a role in overall satisfaction.

The remainder of this paper is organized as follows: Section 2 briefly describes several pre-trained models of deep neural networks for unsupervised text, Section 3 introduces the architecture of the ERNIE-based customer satisfaction analysis model, Section 4 details the dataset construction and annotation principles, and Section 5 analyses the results obtained by our proposed model and two baseline models.

2. Related Work

Recently, pre-trained models of deep neural networks for unsupervised text have significantly improved the performance of multiple NLP (Natural language processing) tasks [5]. Early efforts focused on modeling context-independent word vectors, while later models such as CoVe [6], ELMo [7], and GPT [8] built statement-level semantic representations. The BERT model [9] proposed by Google achieves better results by predicting masked words and exploiting the bidirectional training of Transformer [10], which is an attention mechanism that learns contextual relations between words (or sub-words) in a text.

Whether the models proposed before, such as CoVe, ELMo, GPT and so on, or the more powerful BERT model, their modeling objects mainly focus on the original language signal, and less on the semantic knowledge units. This problem is especially pronounced in Chinese. It is challenging to learn the complete semantic representation of larger semantic units when BERT models Chinese language by predicting Chinese words. For instance, BERT can easily infer the information of masked words by word collocation but cannot explicitly model the semantic units and their corresponding semantic relations [11].

Conceivably, if the model can learn the potential underlying knowledge in massive text, it is bound to further enhance the efficiency of various NLP tasks. The knowledge-enhanced ERNIE model proposed by Baidu learns real-world semantic relations by modeling the prior knowledge such as entity concepts in massive data [4]. Specifically, ERNIE learns the semantic representation of complete concepts by masking semantic units such as words and entities. Compared to BERT, which learns the raw language signals, ERNIE directly models the prior knowledge units and enhances the semantic representation capability of the model [12].

3. Approach

Our model takes a pre-trained ERNIE model as its core, as shown in Figure 1. An input sentence is tokenized into a sequence of tokens. The generated tokens are then put through ERNIE, an enhanced
language representation model which can leverage lexical, syntactic and knowledge information simultaneously. Next, ERNIE yields an embedding sequence \((e_1, e_2, \ldots, e_n, h_c)\) of length \(n + 1\) (the final token \(h_c\) denotes a specific classifier token that captures the overall sentence context).

Figure 1. The architecture of the ERNIE-based customer satisfaction analysis model.

As a text classification task, the representation \(h_c\) of the [CLS] is a fixed dimensional pooled representation of the sequence of tokens, which can be regarded as the sentence-level feature since the context constitutes an important source of disambiguation. This representation \(h_c\) is then fed into a softmax classifier:

\[
\hat{y} = \text{softmax}(W \cdot h_c + b),
\]

Which yields a posterior for each class, i.e., satisfaction and unsatisfaction.

4. Dataset

4.1. Data Collection
The source of the data to be annotated is the transcribed text of the conversation between the customer and the customer service representative and meets the following requirements.

1) No less than 3 effective customer statements per conversation, with no less than 10 words per effective statement.
2) The text recognition accuracy of voice transcription is at least 75%.
3) The call is the incoming call to customer service hotline, without special requirements for province, time, customer attributes, etc.
4) To ensure the accuracy of the model training, the number of annotated data for each perspective is not less than 1000.

4.2. Data Properties
The dataset depicts four perspectives of customer service satisfaction, and their corresponding definitions are shown below.

- **Overall**: Starting from discovering the unsatisfaction, if the customer's expression turns out to be highly positive after the conversation, it can be annotated as satisfied.
- **Product**: Evaluating whether the customer is unsatisfied or deeply confused with the product information, handling, and promotions, etc.
- **Service**: Evaluating whether the customer is unsatisfied with the service skill and attitude of the representative, such as poor attitude, impatience, indifference, etc.
- **Process**: Evaluating whether the customer is unsatisfied with the handling process, such as complaints about identification, forwarding, etc.
4.3. Annotation Principles
To ensure that an accurate and effective model can be trained with the annotated data, we adhere to the following principles in the annotation process.

1) **Discard business experience**: Customer satisfaction metrics need to be evaluated from the consumer's viewpoint. Therefore, business experience should be discarded in the annotation process.

2) **Combine contextual semantics**: Since redundancies, errors, missing words occur frequently in the process of converting conversation to text, it is difficult to determine the real intent of the customer with just one sentence. In this case, contextual semantics should be considered to decide customer satisfaction.

3) **Evaluate based on text only**: For ensuring the validity of the training data, the annotated result can only be evaluated by the textual content. That is, we cannot do any association, subjective guess, or empirical judgment beyond the text.

4.4. Data Processing
We annotate about 1500 entries of data and scored each dialogue text from four perspectives: product, service, process and overall, where the label is marked as satisfied or unsatisfied. Notably, the dialogue text transcribed from customer service consultation is interspersed with noisy data, such as dialect, stutter, lisp, etc., and the scale of the existing annotated data is not so large. Therefore, given the large noise and insufficient amount of the existing data, we expand the data from around 1500 entries to around 14000 entries by EDA (easy data augmentation) [13], which is 9 times the original size.

Furthermore, since the four perspectives of product, service, process, and overall represent different satisfaction scenarios, each perspective in the corpus has a distinct proportion of positive and negative samples. Thus, we manually construct a test set with a 2-to-1 ratio of positive to negative samples and a training set containing all entries except the test set. By doing so, we ensure that our trained model can perform effectively. In this case, the positive represents satisfaction while the negative represents unsatisfaction.

5. Experiments
5.1. Baselines
We compare our model with the following baseline models:

- **BERT-base**: As a baseline, the representation of [CLS] token obtained from the BERT model is treated as the input of the MLP module [14]. The softmax function is then applied to the output of the MLP module to predict the label of customer satisfaction.

- **BERT+DPCNN**: DPCNN is a CNN-based model using word-level deep pyramid, which can effectively extract the remote relation features in text [15]. Like the BERT-base mentioned above, after obtaining the representation of the [CLS] token from the BERT model, it is fed into the DPCNN model with the softmax function attached behind to see whether the customer is satisfied with the service.

5.2. Experiment Settings
We use the ERNIE-base model with a hidden size of 768, 12 Transformer blocks and 12 self-attention heads as well as BERT. We further train the ERNIE-base model on 1 GTX 1080 Ti GPU and set the batch size to 16 to ensure that the GPU memory is fully utilized, with max sequence length of 500. The dropout probability is always kept at 0.1. We use Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We use slanted triangular learning rates, the base learning rate is $2e^{-6}$, and the warm-up proportion is 0.05. We empirically set the max number of the epoch to 10 and the training will be ended prematurely if the performance is not enhanced after training more than 3000 batches.
5.3. Metrics
We use three metrics to evaluate our model: accuracy, recall and F1 score. For testing phase, the result of the sample can be one of true positive (TP), true negative (TN), false negative (FN) and false positive (FP). Mathematically, accuracy, recall and F1 score are defined as follows.

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{3}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{4}
\]

\[
F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}
\]

| Model       | Acc   | Service | Product | Process | Overall |
|-------------|-------|---------|---------|---------|---------|
| BERT-base   | 0.815 | 0.74    | 0.715   | 0.765   |
| Recall      | 0.8833| 0.9173  | 0.8373  | 0.95    |
| F1          | 0.8957| 0.8243  | 0.8299  | 0.8661  |
| BERT+DPCNN  |       |         |         |         |
| Recall      | 0.9   | 0.725   | 0.83    | 0.8     |
| F1          | 0.9474| 0.8421  | 1.0     | 1.0     |
| Ours        |       |         |         |         |
| Recall      | 0.9833| 0.9023  | 0.9880  | 0.975   |
| F1          | 0.9516| 0.8451  | 0.9266  | 0.9150  |

5.4. Results
Since the expression ability of customers varies and the speech recognition of consultations is not quite perfect, the generalization ability of the model is essential. We test our model as well as other two baseline models from four perspectives: product, service, process, and overall, on test set, each of which contains 200 annotated dialog texts, and compared their performance in terms of accuracy, recall, and F1 score.

As depicted in Table 1, for all four perspectives, our model significantly outperforms all baseline models in terms of accuracy and F1 score. Specifically, our model achieves an improvement of at least 2.1\% F1 score compared to BERT and at least 0.4\% F1 score compared to BERT+DPCNN. In addition, for service, process and overall, our model achieves at least 10.0\%, 15.1\% and 2.6\% recall improvement compared to the BERT, respectively.

For different perspectives, the model has different performance in terms of accuracy, recall and F1 score. As an example, our model reaches over 90\% F1 value in overall, service and process, but only reaches 84.5\% in product, indicating that the differences between the metrics of customer satisfaction are quite significant. We believe that such a situation occurs potentially due to the small proportion of negative samples in the training set corresponding to product.

Remarkably, in the perspective of overall, service and process, the recall of BERT+DPCNN on the test set is 1.0, the theoretical maximum, which indicates that BERT+DPCNN fails to learn the knowledge related to customer service satisfaction. This proves that the original BERT is already effective enough and feeding BERT as an embedding layer into other models will reduce the effect of the model instead.

In conclusion, given that BERT+DPCNN cannot learn customer satisfaction-related knowledge well in train sets with disparate proportions of positive and negative samples, our method performs the best results and stability compared to two baseline models, which demonstrates that our method has certain generalization ability.
Table 2. Evaluation results on our model.

| Case Type | Product | Service | Process | Overall | Case Rate |
|-----------|---------|---------|---------|---------|-----------|
| 1         | X       | O       | O       | X       | 0.0332    |
| 2         | O       | X       | O       | X       | 0.1852    |
| 3         | O       | O       | X       | X       | 0.0026    |
| 4         | O       | O       | O       | X       | 0.0114    |
| 5         | X       | X       | O       | X       | 0.2682    |
| 6         | X       | O       | X       | X       | 0.0128    |
| 7         | O       | X       | X       | X       | 0.0830    |
| 8         | X       | X       | X       | X       | 0.4036    |

Perspective Rate: 0.7178, 0.94, 0.502, 1.0, 1.0

Note: 'X' means that the model marks the perspective as unsatisfied, while 'O' means satisfied. Rate stands for the proportion of cases of this type among all cases with overall unsatisfaction, and All stands for the proportion of cases with corresponding unsatisfied perspective among all cases with unsatisfied overall perspective.

5.5. Analysis
To have a deeper insight into how to improve customer satisfaction, we extract 9,000 entries of data from actual customer service logs and evaluate them using our previously trained model. Since the most important customer satisfaction metric is overall, we pick all the data with overall unsatisfaction for analysis.

As shown in Table 2, the proportion of unsatisfied product, unsatisfied service and unsatisfied process in the overall unsatisfied cases are 0.7178, 0.94 and 0.502 respectively (last row), which shows that if the overall satisfaction of a case is unsatisfied, it is extremely likely that the service satisfaction is also unsatisfied, followed by the product satisfaction and the process satisfaction. On the contrary, if a company can pay great attention to the service attitude of customer service representatives, then it is very likely to gain a high level of overall satisfaction.

Besides, the proportion of No. 8, No. 5, and No. 2 cases in the overall unsatisfied cases are 0.4036, 0.2682 and 0.1852 respectively (the last column). Significantly, No. 2 case contains only service unsatisfaction, No. 5 case contains product unsatisfaction on top of that, while No. 8 case contains unsatisfaction from all perspectives. It is evident that service has far more influence on the overall satisfaction than product and process.

In summary, service satisfaction is the most vital part of customer satisfaction, which suggests that overall satisfaction will be enhanced significantly if service satisfaction can be improved. Besides, product and process also have a role in overall satisfaction.

6. Conclusion
We introduce the ERNIE-based customer satisfaction analysis model, treating the task of evaluating customer satisfaction as a text classification task. Our model can automatically parse the text of conversations between the customer and the customer service representative from four perspectives (viz. product, service, process and overall). Further, we construct a corpus containing around 1,500 entries of dialogues texts transcribed from customer service consultation for training our model. Experimental results show that, our model outperforms all baseline models and achieves great generalization capabilities. We also find that among the three perspectives of customer satisfaction (excluding overall), service satisfaction is the most critical. In the future, we will probe more insight of ERNIE on how to improve text representation.

Acknowledgments
The research work is supported by National Key Research and Development Program of China under Grant No. 2018YFC0830705.
References

[1] Daniel E Innis and Bernard J La Londe. Customer service: the key to customer satisfaction, customer loyalty, and market share. *Journal of business Logistics*, 15(1):1, 1994.

[2] Martin Dresner and Kefeng Xu. Customer service, customer satisfaction, and corporate perfo. *Journal of Business logistics*, 16(1):23, 1995.

[3] Amresh Kumar and Bhawna Anjaly. How to measure post-purchase customer experience in online retailing? a scale development study. *International Journal of Retail & Distribution Management*, 2017.

[4] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*, 2019.

[5] Yifan Zhou. A review of text classification based on deep learning. In *Proceedings of the 2020 3rd International Conference on Geoinformatics and Data Analysis*, pages 132–136, 2020.

[6] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. In *Advances in neural information processing systems*, pages 6294–6305, 2017.

[7] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237, 2018.

[8] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, Volume 1 (Long and Short Papers), pages 4171–4186, 2019.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[11] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer, 2019.

[12] Nina Poerner, Ulli Waltinger, and Hinrich Sch¨ utze. Bert is not a knowledge base (yet): Factual knowledge vs. name-based reasoning in unsupervised qa. *arXiv*, pages arXiv–1911, 2019.

[13] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. 2019.

[14] Shaomin Zheng and Meng Yang. A new method of improving bert for text classification. In *International Conference on Intelligent Science and Big Data Engineering*, pages 442–452. Springer, 2019.

[15] Rie Johnson and Tong Zhang. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 562–570, 2017.