

Measurement and Identification of Informative Reviews for Automated Summarization

1st Huyen Nguyen

*dept. of Information Science
University of North Texas
Denton, Texas, USA
HuyenNguyen5@my.unt.edu*

2nd Haihua Chen

*dept. of Information Science
University of North Texas
Denton, Texas, USA
haihua.chen@unt.edu*

3rd Roopesh Maganti

*dept. of Information Science
University of North Texas
Denton, Texas, USA
RoopeshMaganti@my.unt.edu*

4th KSM Tozammel Hossain
*dept. of Information Science
University of North Texas
Denton, Texas, USA
Tozammel.Hossain@unt.edu*

5th Junhua Ding
*dept. of Information Science
University of North Texas
Denton, Texas, USA
junhua.ding@unt.edu*

Abstract—This research investigates the impact of data quality to the quality of text summarization using the software review summarization as a case study. It answers three research questions: 1. What is the most important quality dimension for measuring the quality of software reviews for fitting the review summarization purpose? Our answer is the informativeness of reviews. We propose a metric to measure informativeness and use it to identify highly informative reviews for training review summarization models. 2. How does the review quality affect the quality of review summarization? We conducted the review summarization experiments with a group of datasets that have different quality settings to answer the question. Based on the experiment results, we propose a sampling method for identifying high quality reviews, and the experiment results indicate that the method can significantly improve quality of review summarization on two large review datasets. Furthermore, the results show that the models trained on the selected dataset maintain a balance of bias and variance. 3. Do all text summarization models perform equally well on the datasets? We conduct a comparative study of review summarization on two state-of-the-art deep learning models BART and T5 to answer the question. The research results showing that identifying highly informative reviews is a new direction for improving quality of review summarization.

Index Terms—Text summarization, abstractive summarization, information quality, software review, informativeness, generative language model

I. INTRODUCTION

Text summarization aims to generate a summary that contains the valuable information for a given collection of text documents [1]. The inputs documents could be news articles, search engine results, scientific articles, emails, social media posts, online reviews, or other texts. Based on the number of input documents, text summarization is divided into single-document summarization and multi-document summarization [1]. Multi-document summarization is more challenging due to the information inconsistency among the multiple documents [2]. Product review summarization is one of the applications of multi-document summarization. It has a direct impact on E-commerce performance and consumer behaviors. Its inputs

are customer-generated reviews that describe their assessment of a product from e-commercial sites such as Amazon, Uber, Yelp, and others, and the expected output is a short summary of these reviews [3]. For instance, studies have applied different techniques to generate summaries of online consumer reviews from different perspectives [4]–[7]. The focus of this research is on the software review summarization.

A software review summarization summarizes the pros and cons of a review target, may contains the description of the major functionality, copyright, customer service, cost and other information [8]. The current software marketplaces such as G2.com¹, Software Advice², GetApp³, Capterra⁴, and others provide general information of the software, the list of user reviews, and functions such as feature ranking and comparison with other similar software. Although some researchers have published results on software review summarization such as Uddin et al. [9] and Yang et al. [10] conducted research on automatic summarization of API Reviews, few studies have attempted to generate a summary based on specific input setting such as producing a review summary of all of the incentivized reviews, whose reviewers received the payment for writing the reviews.

There are two types of summarization techniques: abstractive summarization, and extractive summarization. Abstractive summarization generates key points from the input text using machine learning and natural language processing tools and algorithms. Extract summarization identifies and extracts key sentences and phrases from the input text to produce a summary of the original text. Different algorithms have been developed for the automated abstractive summarization [1], [11]–[13], while other studies aim to identify additional features [14], [15], external resources [16], and high-quality

¹<https://www.g2.com>

²<https://www.softwareadvice.com>

³<https://www.getapp.com>

⁴<https://www.capterra.com>

reviews [5], [6] to enhance the text summarization. This study focuses on abstractive summarization.

Both the summarization models/algorithms and the data for training the models/algorithms affect the quality of the software review summarization. The computing rule of “garbage in, garbage out” is still applicable to text summarization [17]. How does the information quality or data quality of the training data impact the quality of text summarization is rarely studied. We use information quality and data quality interchangeably in this article. Therefore, this research aims to examine how the data quality of reviews will affect the quality of the software review summarization. As defined by Chen et al. [17], data quality is characterized by multiple quality dimensions such as credibility, informativeness, readability, and helpfulness, which can be used for measuring the review quality [5], [18]–[20]. Other dimensions include objectivity, relevancy, timeliness, completeness, appropriate amount of information, and ease of understanding [4]. Among all the quality dimensions, informativeness is the most critical dimension for software review summarization. Therefore, we propose and compare three strategies for identifying informative reviews for the review summarization in this research.

In summary, this study will answer three questions regarding evaluation and identification of informative reviews for producing high quality software review summarization.

- 1) How do we quantitatively measure the informativeness of software reviews? Can state-of-the-art prompt-based models such as ChatGPT be used to measure and identify informative reviews? In this paper, we explore GPT3 prompt-based model and propose our metric for measuring and identifying highly informative reviews.
- 2) Are the pre-trained deep learning models generally effective for software review summarization? If they do, which model is the best? Deep learning models have been proven more effective than traditional machine learning models in many studies [21], especially the recent advances in pre-trained deep learning models. Transformers have motivated a corresponding shift to pre-trained methods in text summarization [22]. We compared the effectiveness of pre-trained transformers BART and T5 in the text summarization task.
- 3) How does the data quality of reviews affect the quality of software review summarization? How do we evaluate the quality of a generated summary? We conduct a group of experiments with different data quality and model settings, and report the measurement of summarization quality on two metrics ROUGE .

II. METHODOLOGY

In this section, we formulate the problem, present different methods to measure informativeness in reviews, and describe the summarization models.

Let $D = (p, r_1, r_k, \dots, r_k)$ be a product instance, where p is the product description, and all r_i are associated reviews. Given a set of product instances $\mathbf{D} = \{D_1, D_2, \dots, D_n\}$, we aim to identify informative reviews in \mathbf{D} . We employ

three strategies to identify these informative reviews and create curated datasets.

An overview of our proposed framework is illustrated in Fig. 1.

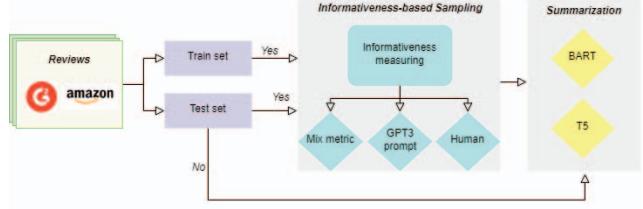


Fig. 1. The proposed framework for identifying high-quality, informative comments. The framework has three modules: a) data extraction and curation, b) informativeness-based sampling, and c) summarizer.

The following subsections describe the informativeness-based sampling method and summarizer.

A. Review informativeness measurement

An informative review is defined as whether the summary conveys important information from the reviews [15], [16]. We expand this notion of informativeness for product reviews and associated review summaries: (1) the informativeness of reviews focuses on whether they contain essential information about the product from product descriptions, and (2) the informativeness of review summaries is determined if they cover salient information from reviews.

We propose a linguistic-based method for identifying informative reviews and compare them against ChatGPT, the state-of-the-art prompt-based model.

1) *Linguistic-mixed method*: The proposed linguistic-based method combines three methods: TF-IDF, named entity recognition (NER), and keyword extraction. These three methods are widely-used to extract key information from the document. TF-IDF quantifies the importance of a word in the document as follows:

$$\text{TF-IDF}_{i,j} = \text{TF}_{i,j} \log \frac{N}{\text{DF}_i} \quad (1)$$

where $\text{TF}_{i,j}$ denotes the number of occurrences of term i in document j , DF_i is the number of documents containing the term i , and N is the total number of documents. We use the trained NER model in SpaCy library to extract NER terms. We extract the keywords using the part-of-speech tagging model built in SpaCy. The words tagged as pronoun (PRON), determiner(DET), and adposition(ADP) are excluded. We then estimate the informativeness using these three linguistic features as follows:

$$I_{LING} = \frac{\alpha|w_{TF-IDF}| + \beta|w_{NER}| + \lambda|w_{KEYWORD}|}{N} \quad (2)$$

where α , β , and λ are hyperparameters, and N is the number of tokens in the review or the summary.

We also assume sentiments expressed in a review contribute to the informativeness of reviews and review summaries. Therefore, we measure sentiment strength as an additional feature in the metric. We use the TextBob library to obtain the sentiment polarity score, which ranges from -1 to 1 . We normalize it by taking its absolute value to get the sentiment strength $I_{SENT} \in [0, 1]$:

$$I_{SENT} = ||S_{sentiment}|| \quad (3)$$

where $||$ denotes the absolute value.

To estimate the relevancy between sequences, we use the ROUGE score [23], the most widely-used metric for evaluating the informativeness of generated summaries w.r.t. the reference. ROUGE-N and ROUGE-L are more commonly-used among all variants of this metric. ROUGE-N measures the number of matching n-grams, while ROUGE-L depends on the longest common subsequence between the two sequences, so having smaller scores. To avoid the vanishing issue of too small ROUGE-L scores, we use ROUGE-N rather. The final informativeness metric is formalized as follows:

$$I_{mix} = \zeta I_{LING} + \gamma I_{SENT} + \tau I_{ROUGE} \quad (4)$$

where ζ , γ and τ are hyperparameters.

2) *Prompt-based method*: This study uses the powerful prompt-based model of OpenAI, *text-davinci-002* [24], to rate the informativeness of reviews and review summaries. The model is based on GPT-3, a large autoregressive language model for language understanding and generation. We define a prompt to rate the informativeness of the target sequence from 0 to 5, given a specific context sequence, similar to our proposed metric. For example, we ask GPT-3 to return the informativeness score for review summaries given product reviews.

B. Abstractive summarization

This study explores the two state-of-the-art Transformers-based sequence-to-sequence models—BART [25] and T5 [26]—for abstractive summarization. BART is a denoising autoencoder seq2seq pretraining for natural language generation, translation and comprehension; it indicatively outperforms existing models on the abstractive summarization task. T5 is also a Transformers model which converts all NLP problems into a text-to-text format. The model is pretrained on multi-tasks with both unsupervised and supervised manners.

III. EXPERIMENT DESIGN AND SETTINGS

For the training and evaluation of our proposed approach, we use two real-world review datasets.

A. Datasets

We collect review data from the world's largest and most trusted software marketplace G2.com⁵. G2 hosts 2,091,100+ authentic, timely reviews from users. We develop a web scraper to extract information of 549 software products along with their reviews and short summaries. This dataset comprises 39,140 product reviews. We use reviews and corresponding summaries as source inputs and reference outputs to train summarization models. Product descriptions are also used as context sequences for the informativeness measurement. The reviews and summaries are quite short compared with common summarization datasets, with average lengths of 36.06 tokens and 5.81 tokens, respectively. Fig. 2 illustrates the length distributions of reviews and summaries in the dataset. The majority of summaries are shorter than ten tokens, while a good number of reviews are less than 30 tokens.

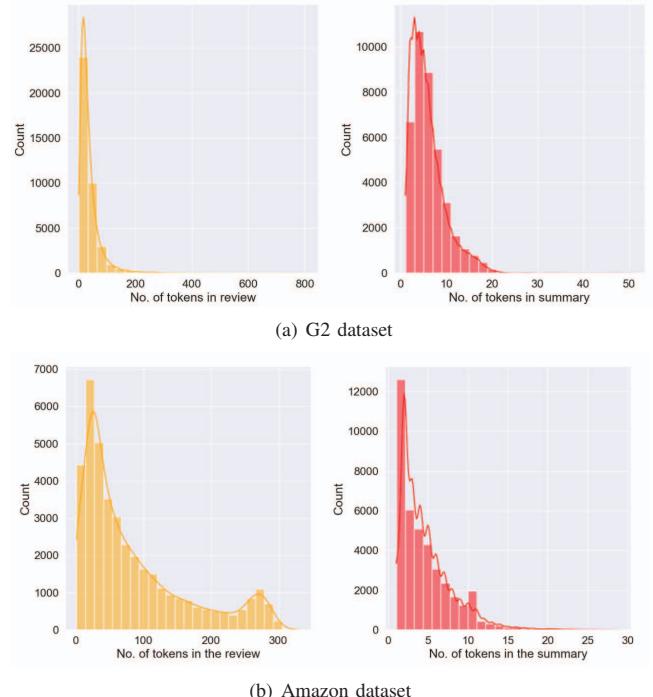


Fig. 2. Lengths of reviews and summaries

We also evaluate our proposed method on the Amazon software review summarization dataset⁶, which we assume to have comparable characteristics such as vocabulary, review, and summary length distributions. Moreover, we attempt to simulate our collected dataset by using an equivalent number of records randomly sampled from the original Amazon software product review summarization dataset (i.e., $n=39,997$ and $n=341,931$ records, respectively). This setup omits the hypothesis that more training data enhance model performance, setting the sampling method as a stand-alone factor that

⁵<https://www.g2.com>

⁶https://huggingface.co/datasets/amazon_us_reviews/viewer/Software_v1_00/train

impacts the performance. Amazon reviews are more lengthy than reviews in G2 dataset, with average lengths of 84.5 tokens and 36.06 tokens, respectively. Meanwhile, the average length of Amazon review summaries is slightly shorter, 4.65 tokens (Fig. 2).

The two datasets are partitioned for training and testing summarization models with a ratio of 0.80 and 0.20, respectively. The sampling methods are applied in each subset independently, as seen in Table I.

B. Informativeness-based sampling

We use informativeness as a data sampling criterion to investigate whether informativeness impacts the summarization. Based on informativeness scores of reviews and review summaries (Figures 3 and 4), we select informativeness thresholds for sampling data.

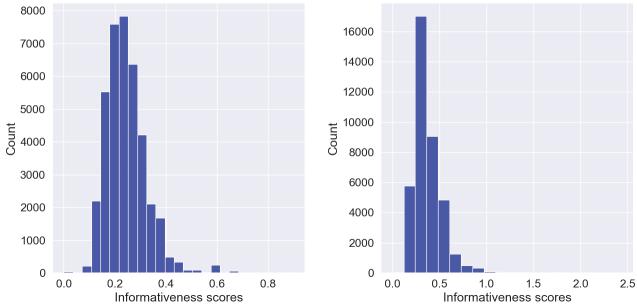


Fig. 3. Distributions of informativeness using our linguistic_mixed approach (train set). Informativeness scores of reviews given product descriptions (left), and informativeness scores of summaries given reviews (right).

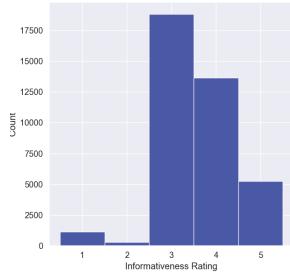


Fig. 4. The distribution of informativeness using the prompt-based method (train set).

Specifically, we create two data samples given by the prompt-based informativeness measurement (called $Info_{prompt}$) with thresholds of 3 and 4. For the informativeness measured by our proposed method (called $Info_{mix}$), we select thresholds of 0.2 or 0.25 to sample data for summarization models. Determining either of these thresholds depends on the datasets we experiment with, as our goal is to use equivalent data sizes among generated samples. With this effort, we aim to remove the assumption of increasing model performance with greater data size.

TABLE I
DESCRIPTION OF THE DATASET WITH OR WITHOUT SAMPLING METHODS.

| Dataset | | Sampling | Sampling on Training | Sampling on Testing | Total |
|---------|-----------------------------|-----------------------|----------------------|---------------------|-------|
| G2 | - | - | ✗ | ✗ | 39140 |
| | W/ sum-based info sampling | $Info_{prompt}^{345}$ | ✓ | ✗ | 38750 |
| | | $Info_{prompt}^{45}$ | ✓ | ✓ | 26456 |
| | | $Info_{mix}^{0.2}$ | ✓ | ✗ | 37540 |
| | | $Info_{mix}^{0.2}$ | ✓ | ✓ | 37144 |
| Amazon | W/ desc-based info sampling | $Info_{mix}^{0.2}$ | ✓ | ✗ | 28554 |
| | | $Info_{mix}^{0.2}$ | ✓ | ✓ | 26055 |
| | W/ sum-based info sampling | - | ✗ | ✗ | 39997 |
| Amazon | | $Info_{mix}^{0.2}$ | ✓ | ✗ | 36654 |
| | | $Info_{mix}^{0.2}$ | ✓ | ✓ | 35818 |

We develop the following data sampling strategies for summarization models:

- $Info_{prompt}^{345}$: This approach samples with prompt-based informativeness rates of 3, 4, and 5.
- $Info_{prompt}^{45}$: This approach samples with prompt-based informativeness rates of 4 and 5.
- $Info_{mix}^{0.2}$: This approach creates samples using our proposed metric with a threshold of 0.2 (only for the G2 dataset).
- $Info_{mix}^{0.25}$: This approach creates samples using our proposed metric with a threshold of 0.25 (only for the Amazon dataset).

We also explore model performances on the test sets with and without informativeness quality control to investigate further noise tolerance of summarization models trained on high informativeness data. The descriptive statistics of the above train samples are presented in Table II.

C. Experiment settings

We intuitively select the same value for all hyperparameters in Eq. 2, $\alpha = \beta = \lambda = 0.3$. Additionally, we choose $\zeta = 1$, $\gamma = 0.1$, $\tau = 2.0$ (Eq. 4). Since ROUGE metric has been widely used in extracting the most important sentences from documents to form summaries, we assign a large value for this hyperparameter (τ).

Our summarization models are initialized from the pre-trained Transformers *BARTlarge*⁷ and *T5large*⁸. All the models are fine-tuned over 20 epochs with the early stopping setting. We use a batch size of 2 for the two models. Adam optimizer with a learning rate of $1e - 4$ is used to train the models. We use a beam search of 3. The experiments are conducted using 2 NVIDIA V100 GPUs.

IV. RESULTS AND DISCUSSION

In this section, we present an evaluation of our proposed approach. We address the following questions:

- 1) RQ1: Does the proposed informativeness measurement metric align with the prompt-based method?
- 2) RQ2: How does the informativeness-based sampling impact the summarization performance? How does our proposed metric compare with the zero-shot GPT3 prompt model?

- 3) RQ3: Can the summarization trained on the high informativeness data tolerate the low informativeness quality?
 4) RQ4: Can our sampling method be generalized to other datasets with similar data attributes?

A. Informativeness measurement

1) *Relation between informativeness measurement and prompt-based method (RQ1):* We measure the correlation between informativeness scores proposed by our method and the zero-shot GPT3 prompt model using Kendall's coefficient of rank correlation (τ_b). Kendall's τ_b is a non-parametric test used to measure the correlation between ordinal and continuous variables [27]. Our proposed metric returns the continuous values, while the prompt-based model provides the ordinal. The analysis shows that the correlation score of the informativeness scores between the two measurements is 0.058, indicating almost no correlation. We will perform additional analysis to understand the issue in the future.

B. Summarization

Table II presents the summarization models' performance on the ROUGE evaluation metric. Overall, ROUGE scores on all models are quite low, suggesting the difficulty of the product review summarization problem. As seen in Figure 2, reference summaries are too short (less than 10 tokens) and highly abstractive, making the summarization more challenging. We found that ROUGE-L scores are quite comparable with ROUGE-1. ROUGE-L considers the ratio of the longest common subsequences between the model's output and the reference, while ROUGE-1 takes single-grams overlapping between both into account.

1) *Comparison between sampling methods (RQ2):* The results show that T5 achieves better performance among the two baseline models, with roughly 1.3% on the ROUGE metric. Therefore, we select T5 to experiment with our proposed sampling methods. As described in Section III, we explore the GPT-3 prompt-based model and then propose a new metric to rate the informativeness of reviews and review summaries. Table II presents models' performance on different sampling methods, including GPT-3 prompt-based (called $Info_{prompt}$), and our linguistics-based method called $Info_{mix0.2}$. The results indicate that the models trained on the data sampled by our informativeness strategy outperform those trained on the data samples created using the prompt-based model by 1-2%, demonstrating the effectiveness of our data sampling method and our proposed informativeness measurement metric.

2) *Noise tolerance of sampling methods (RQ3):* Assuming that sampling both train and test sets with the same method results in similar data distributions, likely leading to satisfactory performance on the test set. Therefore, we also investigate the model performance on the test set, which is not applied any sampling methods. The experiment results reveal that the performance is equivalent to or slightly lower (around 0.1%) than the one tested with the sampled data. It implies that this sampling method does not negatively affect the model's noise tolerance.

TABLE II
 PERFORMANCES OF SUMMARIZATION MODELS ON G2 DATASET. RESULTS ARE REPORTED ON THE TEST SET.

| | Model | Sampling | R-1 | R-2 | R-L |
|------------------------------------|-----------------------------|----------------------|---------------|--------------|---------------|
| Baseline | <i>BART_{large}</i> | - | 10.341 | 1.200 | 9.536 |
| | <i>T5_{large}</i> | - | 11.308 | 3.001 | 10.978 |
| W/ sum-based info sampling | <i>T5_{large}</i> | <i>S₁</i> | 10.809 | 2.739 | 10.580 |
| | <i>T5_{large}</i> | <i>S₂</i> | 10.151 | 2.562 | 9.742 |
| | <i>T5_{large}</i> | <i>S₃</i> | 12.107 | 3.489 | 11.752 |
| | <i>T5_{large}</i> | <i>S₄</i> | 10.642 | 2.681 | 10.408 |
| | <i>T5_{large}</i> | <i>S₅</i> | 10.015 | 2.515 | 9.606 |
| | <i>T5_{large}</i> | <i>S₆</i> | 11.867 | 3.398 | 11.529 |
| W/ desc-based info sampling | <i>T5_{large}</i> | <i>S₃</i> | 12.562 | 1.577 | 9.577 |
| | <i>T5_{large}</i> | <i>S₆</i> | 12.688 | 1.570 | 9.637 |

Note: $S_1 = Info_{prompt}^{345\ train+test}$, $S_2 = Info_{prompt}^{45\ train+test}$
 $S_3 = Info_{ling}^{0.2\ train+test}$, $S_4 = Info_{prompt}^{345\ train}$
 $S_5 = Info_{prompt}^{45\ train}$, $S_6 = Info_{ling}^{0.2\ train}$

TABLE III
 PERFORMANCES OF SUMMARIZATION MODELS ON AMAZON DATASET.
 RESULTS ARE REPORTED ON THE TEST SET.

| | Model | Sampling | R-1 | R-2 | R-L |
|-----------------------------------|-----------------------------|----------------------|---------------|---------------|---------------|
| Baseline | <i>BART_{large}</i> | - | 7.177 | 0.669 | 6.859 |
| | <i>T5_{large}</i> | - | 8.295 | 3.168 | 7.697 |
| W/ sum-based info sampling | <i>T5_{large}</i> | <i>S₇</i> | 9.281 | 3.919 | 8.741 |
| | <i>T5_{large}</i> | <i>S₈</i> | 16.544 | 10.791 | 16.179 |

Note: $S_7 = Info_{ling}^{0.25\ train+test}$, $S_8 = Info_{ling}^{0.25\ train}$

Surprisingly, the $Info_{prompt}$ models consistently perform worse than the baseline (by almost 1%). To our knowledge, the prompt-based model we use is among the state-of-the-art and is trusted to use as a powerful zero-shot model. It raises a big concern about the accuracy of this model for the informativeness rating problem. The size of the dataset sampled by this informativeness model is even smaller than the one created with our method (by about 4%). That means the worse performance of models trained on $Info_{prompt}$ data is not caused by less training data.

3) *Sampling method performance on Amazon dataset (RQ4):* We further test our proposed informativeness-based sampling method on the Amazon software product review dataset. The results are presented in Table III. Based on the results of our baseline testing, T5 continues to outperform BART, suggesting that we should explore our sampling method further with the T5 model. Overall, the T5 baseline performs worse on this dataset than on the G2 dataset. As aforementioned, we randomly sampled the Amazon dataset to create the data with a comparable size to our collected dataset (Table I). Additionally, we selected an appropriate informativeness threshold so that the sampled data has a similar size to the G2 informativeness-sampled data (See Table I). Again, the similar training data size can reject the hypothesis that the performance improvement is led by having more training data, making it easier to conclude the effectiveness of our method.

The T5 model, trained on the data sampled with our

informativeness method, outperforms the baseline with almost double ROUGE scores; it even exceeds the similar model on the G2 dataset. The model tested on the test set without the sampling method performs much better than on the sampled test set, consistently concluding our sampling method does not reduce the noise tolerance capacity of the model.

V. CONCLUSION

This study proposes a metric to quantitatively measure the informativeness of software product reviews and review summaries. We compare our metric with the zero-shot GPT3 prompt-based model on measuring the informativeness of reviews and associated summaries. We enhance state-of-the-art summarization models using the obtained informativeness levels to sample the data for training the models. Our experiment results indicate that our sampling method is significantly effective, improving summarization by a large margin on both G2 and Amazon datasets. Further, we investigate how the summarization models trained on high-informative inputs perform on the data with low informativeness quality. The results suggest that the model performance does not drop, demonstrating that the models trained on data selected with this sampling method do not suffer from overfitting; in contrast, it can maintain a good balance of bias and variance. In the future, we would like to compare our automatic method with human evaluation to measure the informativeness of review contents and review summaries. We will also implement other metrics for software review summarization, such as BERTscore.

REFERENCES

- [1] M. Gambhir and V. Gupta, "Recent automatic text summarization techniques: a survey," *Artificial Intelligence Review*, vol. 47, pp. 1–66, 2017.
- [2] L. Huang, Y. He, F. Wei, and W. Li, "Modeling document summarization as multi-objective optimization," in *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*. IEEE, 2010, pp. 382–386.
- [3] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '04. New York, NY, USA: Association for Computing Machinery, 2004, p. 168–177. [Online]. Available: <https://doi.org/10.1145/1014052.1014073>
- [4] C. C. Chen and Y.-D. Tseng, "Quality evaluation of product reviews using an information quality framework," *Decision Support Systems*, vol. 50, no. 4, pp. 755–768, 2011.
- [5] C.-F. Tsai, K. Chen, Y.-H. Hu, and W.-K. Chen, "Improving text summarization of online hotel reviews with review helpfulness and sentiment," *Tourism Management*, vol. 80, p. 104122, 2020.
- [6] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou, "Low-quality product review detection in opinion summarization," in *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, 2007, pp. 334–342.
- [7] J. Yi and Y. K. Oh, "The informational value of multi-attribute online consumer reviews: A text mining approach," *Journal of Retailing and Consumer Services*, vol. 65, p. 102519, 2022.
- [8] Z. Ni, M. Wölk, G. Jukes, K. Mendivelso Espinosa, R. Ahrends, L. Aimo, J. Alvarez-Jarreta, S. Andrews, R. Andrews, A. Bridge *et al.*, "Guiding the choice of informatics software and tools for lipidomics research applications," *Nature Methods*, vol. 20, no. 2, pp. 193–204, 2023.
- [9] G. Uddin and F. Khomh, "Automatic summarization of api reviews," in *Proceedings of the 32nd IEEE/ACM International Conference on Automated Software Engineering*, ser. ASE '17. IEEE Press, 2017, p. 159–170.
- [10] C. Yang, B. Xu, J. Y. Khan, G. Uddin, D. Han, Z. Yang, and D. Lo, "Aspect-based api review classification: How far can pre-trained transformer model go?" in *2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 2022, pp. 385–395.
- [11] S. Gupta and S. K. Gupta, "Abstractive summarization: An overview of the state of the art," *Expert Systems with Applications*, vol. 121, pp. 49–65, 2019.
- [12] A. P. Widjassari, S. Rustad, G. F. Shidik, E. Noersasongko, A. Syukur, A. Affandy *et al.*, "Review of automatic text summarization techniques & methods," *Journal of King Saud University-Computer and Information Sciences*, 2020.
- [13] C. Ma, W. E. Zhang, M. Guo, H. Wang, and Q. Z. Sheng, "Multi-document summarization via deep learning techniques: A survey," *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–37, 2022.
- [14] S. Urolagin and L. Satish, "Improving the quality of text summarization using pronoun replacement technique," in *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. IEEE, 2017, pp. 1991–1995.
- [15] P. Cui and L. Hu, "Topic-guided abstractive multi-document summarization," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 1463–1472.
- [16] W. Wu, W. Li, X. Xiao, J. Liu, Z. Cao, S. Li, H. Wu, and H. Wang, "Bass: Boosting abstractive summarization with unified semantic graph," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 6052–6067.
- [17] H. Chen, J. Chen, and J. Ding, "Data evaluation and enhancement for quality improvement of machine learning," *IEEE Transactions on Reliability*, vol. 70, no. 2, pp. 831–847, 2021.
- [18] J. Mackiewicz and D. Yeats, "Product review users' perceptions of review quality: The role of credibility, informativeness, and readability," *IEEE Transactions on Professional Communication*, vol. 57, no. 4, pp. 309–324, 2014.
- [19] A. Boluki, J. P. R. Sharami, and D. Shterionov, "Evaluating the effectiveness of pre-trained language models in predicting the helpfulness of online product reviews," *arXiv preprint arXiv:2302.10199*, 2023.
- [20] X. Cai, J. Cebollada, and M. Cortiñas, "Impact of seller-and buyer-created content on product sales in the electronic commerce platform: The role of informativeness, readability, multimedia richness, and extreme valence," *Journal of Retailing and Consumer Services*, vol. 70, p. 103141, 2023.
- [21] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert systems with applications*, vol. 165, p. 113679, 2021.
- [22] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, "Summeval: Re-evaluating summarization evaluation," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 391–409, 2021.
- [23] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.
- [24] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [25] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880.
- [26] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.
- [27] H. Khamis, "Measures of association: how to choose?" *Journal of Diagnostic Medical Sonography*, vol. 24, no. 3, pp. 155–162, 2008.