Hotel Reviews AI Solutions:
Triple Win for Customers, Business, and Platform

Mengting Feng
121090123@link.cuhk.edu.cn

Yumeng Peng
121090444@link.cuhk.edu.cn

Yingxuan Bian
120050052@link.cuhk.edu.cn

Jiarui Zhang
120030058@link.cuhk.edu.cn

Xinyi Huang
119020214@link.cuhk.edu.cn

1 Background

1.1 Business Problems

Reviews are important for both businesses and customers on the online hotel reservation platform. However, they are facing challenges on several aspects.

For hotel businesses:
Businesses have low efficiency and effectiveness in replying to customer reviews.
a. The current generic responses fail to achieve effective interaction with customers.
b. Editing and replying to comments is time-consuming.
c. Some businesses struggle to produce well-crafted responses due to a lack of writing skills.

Businesses find it difficult to extract useful information and generate effective optimization strategies from reviews.
a. The information in the reviews is chaotic, hard to summarize.
b. Reviewing, analyzing, and summarizing reviews, as well as generating optimization strategies, require a substantial amount of time.
c. Some businesses lack the ability to derive actionable insights from evaluations.

For customers:
Customers have difficulties in getting useful information from reviews.
a. The large number of fake reviews are annoying.
b. The large number of reviews contains chaotic information, hard for customers to get useful information.
c. Hard to get overall information containing reviews related to all aspects of the hotel.

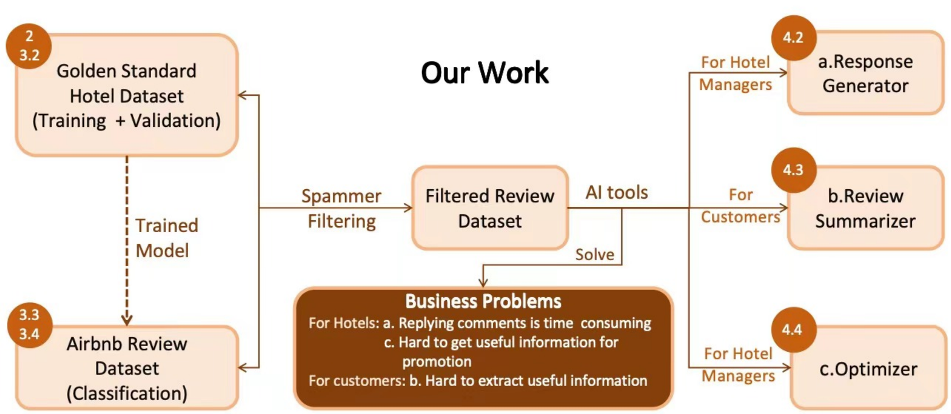
1.2 Motivation

As discussed in the Business Problems part, there are several gaps between the current situation and users' needs. Our motivation is to develop AI-driven solutions bridge these gaps.

The solutions will enhance the efficiency of review management of hotel businesses, helping them to generate customized replies to customers' reviews, summarize the key information of the reviews, and provide them with optimization strategies.

Customers will be provided with clear and comprehensive summary of the hotel reviews, helping them in the decision-making process. Due to the enhanced experiences for both hotels and customers, the retention rate of the hotel booking platform is also expected to increase.

1.3 Introduction of the Whole Solution



Our solution is AI tools that revolutionize the way hotel businesses and customers interact with online reviews. The solution consists of two main parts:

Step1: Spammer Filtering. We employed advanced algorithms to detect spam and fake reviews, providing a reliable foundation for lager part analysis and usage.
Step2: We designed 3 AI tools to address the 3 pain points faced by hotel businesses and customers. Review summarizer helps customers to extract and structure useful information. Response generator can help hotel managers to generate customized replies to the reviews. Report generator summarizes the key data and information for hotel managers, and provides useful suggestions for optimization.

2 Data Pre-processing

In order to solve our business problem and develop useful AI tools, we must first get a clean and valid review set with spammer comments filtered out. We used two datasets for our spammer detection and filtering.

2.1 Dataset Introduction

Firstly we used a golden standard hotel comment dataset for model training and validation. This corpus consists of truthful and deceptive hotel reviews from 20 hotels in the Chicago area. This dataset contains: 400 truthful reviews from TripAdvisor.com and 400 deceptive reviews from Amazon Mechanical Turk [7]. Specifically, TripAdvisor reviews contain only the features (unigrams and bigrams), rather than the original reviews. This is done to protect the privacy of the original posters. Features additionally encode POS information as given by the Stanford Tagger [5]. While Mechanical Turk reviews are given in both feature-only and raw formats.

Then, we collect an Airbnb review dataset from https://insideairbnb.com/get-the-data/ for spammer filtering and AI tools implementation. This data set includes two parts: review and listing. The first part contains the review content information, and the second part is the house information of the corresponding review, including the hotel name, location, picture url, number of bathrooms and other detailed information.

2.2 Data Cleaning and Manipulating

For training data, we read all the tokenized text training data from the folder labeled 'TripAdvisor' and 'MTurk' one by one and convert them to the "train_data" DataFrame format. This DataFrame is meticulously crafted to include crucial details such as the hotel's name ('Comment Name'), the actual review content in both unigram and bigram formats ('Unigram' and 'Bigram'), and the corresponding class label ('Class') based on the source folder ('TripAdvisor' or 'MTurk'), with 1 indicates real review and 0 indicates spammer one.

For testing data, we firstly loaded the hotel name and comment content columns of original Airbnb review dataset, with a fraction of this data sampled for processing, maintaining a balance between computational efficiency and data representation. Following this, we merged the reviews and listings data based on "id". Then, we filtered out non-string values from the 'comments' column, encoded the textual data into unigrams and bigrams using established NLTK tools, and organized them into the same format as the training data.

3 Spammer Filtering

3.1 Literature Review

To accomplish our task of building AI models, the first step is to extract key information from millions of reviews. However, online reviews have relatively low credibility. In 2022, a survey showed that 30%-40% of reviews are fake, and 28% of customers do not trust online reviews [9]. Fake comments mainly consist of two categories: one is merchants hiring fake reviewers to leave positive reviews, and the other is encouraging customers to leave positive reviews in exchange for rewards [9]. These large amounts of fake positive reviews will cause a significant bias to our research. To address this issue, we have decided to build a comment filter to identify real and fair comments.

From previous literature, we found that most studies constructed multivariate distribution models based on user account information and the content of comments. Ott et al. compared three supervised models to detect deceptive reviews, achieving 90% accuracy [8]. Moreover, this experiment provided a golden-standard dataset, which was also used in our experiment. Kumar [6] utilized neural networks to detect the authenticity of comments. Elmogy's research [3] considered incentivized reviews and judged the authenticity of comments based on user features and behaviors.

Learning from previous research, with the consideration that businesses now tend to incentivize customers compared to hiring water armies [4], we have shifted our focus more towards the content of the text itself to filter out real comments.

3.2 Methodology

We incorporated three distinct approaches for training and classification. The first approach involves utilizing Word2Vec embeddings with a Random Forest Classifier. After preprocessing, the tokenized sequences were used to train a Word2Vec model, capturing semantic relationships among words. Subsequently, average word vectors were generated for each comment in both the training and testing datasets, serving as features for training the Random Forest Classifier. The training performance was further evaluated using validation metrics such as accuracy, precision, recall, and F1 score.

In the second approach, the Word2Vec embeddings were trained with a Neuron Network Classifier. Similar preprocessing steps were undertaken. However, instead of a Random Forest Classifier, a deeper neuron network model was constructed using TensorFlow's Keras API. This neuron architecture comprises multiple dense layers with varied neuron counts, enabling the model to learn complex feature representations. The deeper neural network model was compiled with binary cross-entropy loss and accuracy metrics, then trained on the Word2Vec embeddings derived from the training data. The model's performance was subsequently evaluated using the same validation metrics to assess its classification capabilities.

Lastly, we implemented a fusion approach by combining TF-IDF features with Word2Vec embeddings in conjunction with a Random Forest Classifier. This combined feature set integrated TF-IDF vectors, which captured word importance in comments, with the semantic context captured by Word2Vec. The merged features of word vectors and TF-IDF feature were then fed into a Random Forest Classifier for training and evaluation on the training-validation split. The predictive performance was assessed using standard evaluation metrics to determine its effectiveness in handling the integrated feature space.

3.3 Result and Evaluation

The four models we constructed: Random Forest, standard neural network (NN), deep neural network (Deep NN), and wide neural network (Wide NN), performed differently on the dataset. We find the standard neural network model gives best precision, which the random forest model generate achieved highest recall and accuracy. Given our preference to avoid overlooking genuine information, we set a higher cost for false negatives (comments that are true but mistakenly classified as false). Overall, we found that Random Forest yielded the best results. However, there is still space for optimization.

Method	**Random Forest**	**Neural Network**	**Deep NN**	**Wide NN**
Accuracy	0.638	0.625	0.625	0.631
Precision	0.631	0.999	0.885	0.644
Recall	0.663	0.250	0.288	0.588
F1 Score	0.646	0.400	0.434	0.614

Table 1: Performance Metrics of Different Methods

When we applied the term frequency–inverse document frequency (TF-IDF) technique to extract features from the comment text and utilized the previously optimized random forest classifier, we observed a significant improvement in the overall accuracy, recall, precision, and F1 score. The result is showing in Table 2 and 3. Consequently, this model was chosen as the final approach for our project.

	Predicted	
Actual	**True**	**Fake**
True	TP (329)	FP (71)
Fake	FN (49)	TN (351)

Table 2: Confusion Matrix

Metric	**Value**
Accuracy	0.850
Precision	0.868
Recall	0.825
F1 Score	0.846

Table 3: Performance Metrics

4 AI Tools

4.1 General Introduction of 3 Tools

After finishing the spammer filtering process, the true reviews on Airbnb are filtered into the following process to build AI tools. Since three business problems remained to be solved, we respectively designed three AI tools targeting each business problem. For hotels, replying to online reviews is very time-consuming, and sometimes the replies are lack of customization. Therefore, we design a response generator bot for hotel managers to generate specialized consumer responses. Consumers find it hard to extract useful information from previous reviews, such as keywords of service quality, the disadvantages, and location comments. The second AI tool, review summarizer, is for them to summarize historical reviews from other consumers and provide them with useful information. Also, the hotels may face a situation where hotel managers find it hard to get useful information from all the online reviews and receive useful suggestions to improve their services. For this business problem, we designed a review report generator for hotel managers to get feedback from many online reviews. Compared to the review summary for consumers, the review report for managers is a step further as an optimizer because it not only analyzes the context of the reviews but also calculates the mean score, collects the number of positive and negative comments, and checks the sentiment of the reviews before generating the report.

To create responses for the reviews on Airbnb, OpenAI GPT-3.5 Turbo is utilized as an advanced text generation model to generate text based on prompts provided by the user. In this project, GPT-3.5 Turbo generates responses for the hotels and consumers. We also improved the prompt engineering to increase the performance of chat generation and sentiment analysis to improve the prompt quality. Gradio is for creating customized UI interfaces. In this project, Gradio is employed to design a user interface that allows hotel managers to input hotel reviews and view the automatically generated responses from the hotel chatbot.

4.2 Specialized Review Response

The first AI Tool is built to generate specialized responses so consumers can get high-quality hotel replies. To call ChatGPT, we first set up OpenAI API by importing packages, setting the API keys, and initializing the OpenAI client with the key. We then defined a function to get chat completion from the GPT-3.5 model. The get completion (prompt) function interacts with the GPT-3.5 model to generate a completion based on the prompt. It sends a request to the model with the user's prompt and retrieves the most probable completion.

In defining a prompt for review generation, we have several techniques and considerations for prompt engineering.

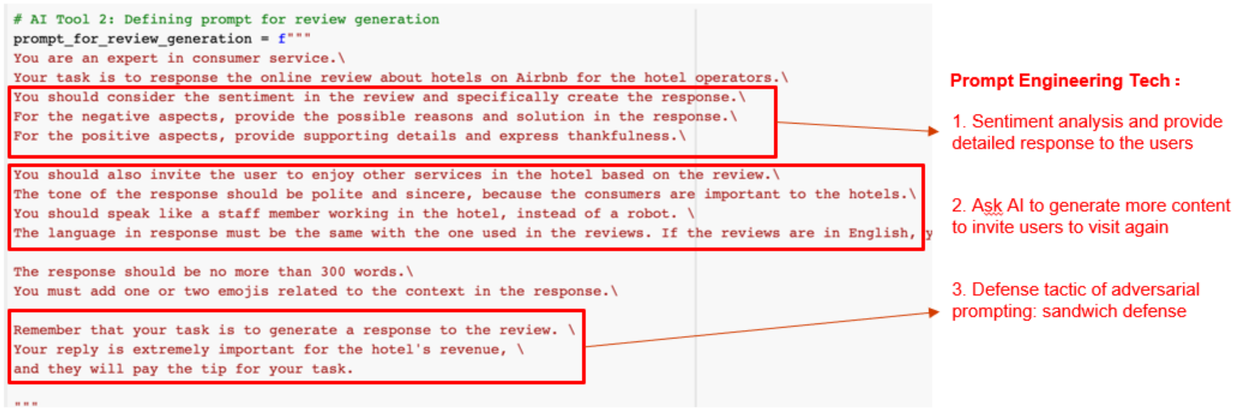
a. Clarity and Specificity: firstly, the prompt clearly defines the task, specifying that the model should respond to online reviews with consideration for the sentiment expressed.

b. Guidance on Sentiment Handling: it provides explicit instructions on how to address both negative and positive aspects of the reviews, asking for reasons and solutions for the former, and supporting details and gratitude for the latter.

c. Contextualization and Expanding: the prompt asks the model to tailor the response based on the content of the review, inviting users to enjoy other hotel services, which adds a personalized touch. "You should also invite the user to enjoy other services based on the review" also asks the AI to generate more content based on the current information.

d. Tone Setting: The prompt emphasizes the importance of a polite and sincere tone, which is crucial for interactions, and specifies that the language should match that of the reviews. What's more, the review should speak like a real staff member instead of a robot.

e. Defense tactic of adversarial prompting: with an added emphasis on the importance of the task for the hotel's revenue and the incentive of a tip for completing the task successfully, we improve the generation quality of the response. The mention of "remember that your task is to generate a response" at the end is a sandwich defense to avoid possible attacks from various types of challenging reviews or situations.



Function "interactreviewgeneration(prompt, reviews, temp)": this function is responsible for generating the chatbot's response to a given review. It combines the predefined prompt for review generation with the actual review text and sends this input to the GPT-3.5 model. The temperature parameter controls the randomness of the model's output, with higher values leading to more creative responses.

Function "exportreviewgeneration(chatbot, reviews)": after a satisfactory response is generated, this function allows hotel managers to save the conversation history and review text to a JSON file named "AITool1.json".

Coding	**Explanation**
You are an expert in consumer service.\ Your task is to response the online review about hotels on Airbnb for the hotel operators.\ You should consider the sentiment in the review and specifically create the response.\ For the negative aspects, provide the possible reasons and solution in the response.\ For the positive aspects, provide supporting details and express thankfulness.\	**1. Sentiment analysis and provide detailed responses to the users**
You should also invite the user to enjoy other services in the hotel based on the review.\	**2. Expanding: ask the AI to generate more content to invite users to visit again**
The tone of the response should be polite and sincere, because the consumers are important to the hotels.\ You should speak like a staff member working in the hotel, instead of a robot. \ The language in response must be the same with the one used in the reviews. If the reviews are in English, your responese should also be in English.\	**3. Tone and language setting**
The response should be no more than 300 words.\ You must add one or two emojis related to the context in the response.\	**3. Word limitation and emojis adding**
Remember that your task is to generate a response to the review. \ Your reply is extremely important for the hotel's revenue, \ and they will pay the tip for your task.	**4. Defense tactic of adversarial prompting: sandwich defense to avoid possible attacks; emotional prompting: tips**

After defining the prompt, we utilized the Gradio library to create a user interface for interacting with the chatbot. Managers can input or modify reviews in the provided text box and adjust the temperature to control the chatbot's response style. Upon clicking the "Send" button, the chatbot generates a response, which is then displayed in the chat interface. If hotel managers are satisfied with the response, they can click the "Export" button to save the conversation.

Result 1: A response to a positive review

We are delighted to hear that you and your boyfriend had a wonderful stay at our charming house by the canal! —The perfect atmosphere and beautiful view of the canal must have been a real treat. We're glad you found the location convenient, as being close to the city center is a top priority for many of our guests.

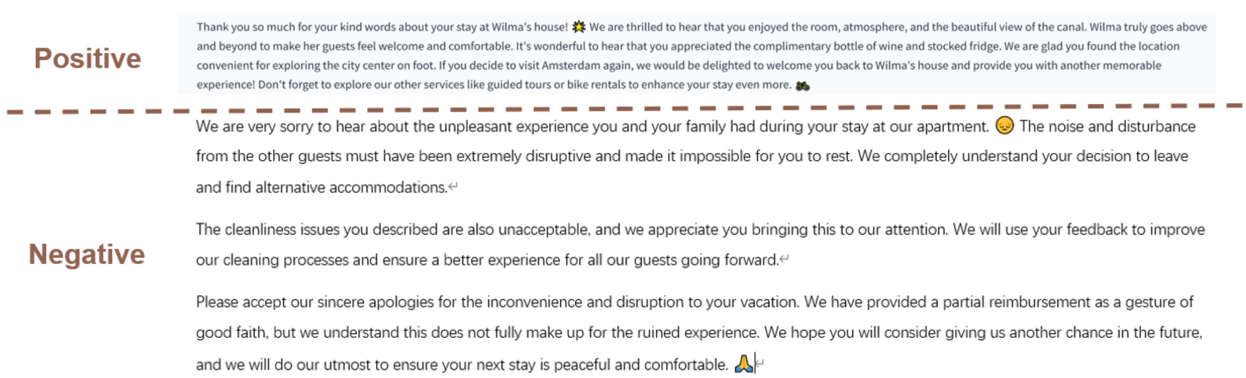
Your kind words about Wilma are very much appreciated. We always strive to provide a warm welcome and the best possible experience for our guests. It's wonderful to hear that the complimentary wine and beverages were a nice touch. —We can't wait to have you back the next time you visit Amsterdam!

Result 2: A response to a negative review

We are very sorry to hear about the unpleasant experience you and your family had during your stay at our apartment. The noise and disturbance from the other guests must have been extremely disruptive and made it impossible for you to rest. We completely understand your decision to leave and find alternative accommodations.

The cleanliness issues you described are also unacceptable, and we appreciate you bringing this to our attention. We will use your feedback to improve our cleaning processes and ensure a better experience for all our guests going forward.

Please accept our sincere apologies for the inconvenience and disruption to your vacation. We hope you will consider giving us another chance in the future, and we will do our utmost to ensure your next stay is peaceful and comfortable.



The first AI tool provides an innovative approach to managing online reviews for hotels on Airbnb. By leveraging the power of the GPT-3.5 model, hotel operators can automate the process of generating thoughtful, personalized responses that reflect the sentiment of each review. This not only saves time but also has the potential to improve guest satisfaction and loyalty.

4.3 Key Information Summarization

For this second tool, our main goal is to utilize AI's summary ability to generate key information from dozens or hundreds of previous reviews, which brings convenience to users.

The prompt engineering techniques we used here are similar to those in the first tool, including Clarity and Specificity, Guidance on Sentiment Handling, and the Defense tactic of adversarial prompting. The implementation steps include three major parts. First, we set up OpenAI API and define the completion function. Considering the max token of GPT-3.5-turbo API, which is 4096 tokens and approximately 3000 words, we randomly selected 30 reviews representitive from the same hotel. Second, we input prompts for three tasks, including a summary of advantages, a summary of disadvantages, and three keywords of features, respectively. Last, we print the response one by one, and a sample response of keywords is as follows:
a. Cleanliness: Rooms and facilities are well-maintained and spotless.

b. Convenient Location: Central near attractions and transportation.
c. Friendly staff: Guests appreciate the helpful and welcoming hotel staff.

4.4 Report Generator

The third tool is provided to hotel managers to help them summarize key information from the reviews and generate optimization suggestions. To get numerical information for the report. We only kept the real comments (predicted class=1) for review summarization. We calculated the number of reviews and mean of scores. Then, we calculated the number of positive and negative reviews after conducting sentiment analysis using TextBlob.

The verbal part of the report is generated by utilizing AI technology. OpenAI API is used to analyze the reviews. Considering the max token of GPT3.5-turbo, we randomly selected 20 reviews from the same hotel as representative reviews. Then, we used the prompt to let the GPT3.5-turbo model summarize the advantages and disadvantages of the hotel and generate optimization suggestions for the disadvantages.

Lastly, the report is generated following a given format. The sample is shown below.

During this period, your hotel has received an average rating of 4.79. You have received a total of 1556 reviews, out of which 1307 were positive and 17 were negative.

Customers have praised the following aspects of your hotel:

1. Excellent location, close to attractions and convenient transportation.
2. Spacious and comfortable rooms with well-equipped bathrooms.
3. Attentive and friendly service with a positive attitude from the staff.
4. A wide variety of breakfast options.
5. The cleanliness and hygiene of the premises have been appreciated by customers.

Customers expressed dissatisfaction with the following:

1. Room cleanliness is inadequate, with issues of dirt and unpleasant odors.
2. Service attitude is not friendly or professional, lacking enthusiasm.
3. Facilities and equipment are significantly outdated and require maintenance.
4. The quality and variety of food at the hotel need improvement.
5. Noise issues affecting guests' rest experience.

You can consider the following optimization suggestions:

1. Inconsistent service quality: Ensure that employees receive professional training to improve overall service levels. Establish clear service standards and processes to ensure that every staff member delivers consistent high-quality service.
2. Delayed facility maintenance: Establish a regular inspection plan for facilities and promptly repair or replace damaged equipment. Invest in facility maintenance and upgrades to ensure guest comfort and safety.
3. Communication issues: Establish effective communication channels, including customer feedback mechanisms and internal communication channels among staff. Encourage employees to provide suggestions and feedback to improve workflow and service quality.
4. Lack of personalized service: Understand guests' needs and preferences to provide personalized service experiences. Train employees on how to better interact with guests to create unique and memorable experiences.
5. Insufficient marketing and promotion: Strengthen marketing activities, both online and offline. Utilize social media platforms and partnerships to increase visibility and attract more customers.

4.5 Evaluation

For generative contents without label, it is not feasable to use predict matrics such as bleu-4, rouge-1, rouge-2 to evaluate. Thus, to assess the effectiveness of three AI tools we build, we conducted experiment by inviting 30 audiences to rate on various corresponding aspects.

For the first tool "Specialized Response Bot", we set four evaluation aspects, which are "Appreciation", "Invite Return", "Reinforce Good" and "Apologize", according to a chart from ReviewTracker. We invite our 30 audiences to rate from 0 to 10. The average score are around 8, with the "Apologize" rating highest at 8.8, and "Invite Return" rating lowest at 7.8. We think it may be because AI is very sensitive to negative sentiment, and is not smart enough to go a step further besides appreciation and reinforcing good.

Evaluation Aspects	Average Rating
· **Thanks and Appreciation**	7.8
· **Reinforce Good**	7.8
· **Apologize**	8.8
· **Invite Return**	9.0

Table 4: Performance Rating for Specialized Response Bot

For the second tool "Key Information Summarization Bot", we conducted the assessment with another four aspects "Friendliness", "Accuracy", "Conciseness", and "Readability". What's more, we compare two versions of summary response, one of which using only basic prompting and another one adding sentimental analysis results in the prompting. The result is that for most of the aspects, two versions of the response do not have significant differences in rating. While on readability, it seems that the version with sentiment analysis is 10 percent lower in rating. The cause behind and the robustness are still worth working on for future research and projects.

Evaluation Aspects	Rating without Sentiment Analysis	Rating with Sentiment Analysis
· **Friendliness**	8.1	7.9
· **Accuracy**	7.8	8.2
· **Conciseness**	9.0	9.2
· **Readability**	9.0	8.0

Table 5: Performance Rating for Key Information Summarization Bot

For the third tool "Report for Business", we keep three aspects "Accuracy", "Conciseness", and "Readability", but substitute "Friendliness" with "Professionality" since it is a more important factor for a report targeting hotel managers. As table 6 shows, the evaluation result was that ratings on the other three aspects are between 8 and 9, however, the rating for "Professionality" is only around 5. It reflects that the professionality of AI-generated reports is a direction.

Evaluation Aspects	Average Rating
· **Professionality**	5.4
· **Accuracy**	8.2
· **Conciseness**	8.8
· **Readability**	8.6

Table 6: Performance Rating for Report Generator Bot

5 Future Work

Though the performance of these AI tools is excellent, there are still some limitations. First, regarding the professionality, we receive comments from audiences that "the report contents are reasonable but as general as common sense". Second, for conciseness, more than 75 percent of audiences think the keywords are "relatively long" or "very long". Third, the tone of AI-generated review responses is not stable. To be more specific, sometimes it sounds too formal, while in other cases it is too passionate.

To tackle these problems, we suggest future researchers and product managers proceed with following directions. Firstly, people can try fine-tuning the bots with industry data such as hotel or tourism, which could enhance the professionality of the report. Secondly, for conciseness issues, it is suggested to add more prompts and token or word limits. Also, it is also worth adopting text summarization [2] and other NLP algorithms and techniques [1] that industry and researchers are working on. Thirdly, in terms of tone issues, human labeling is a feasible way to assess the quality. Then, with increasing labeled AI-generated datasets for future model training, AI tools may perform better with a more humanlike tone.

6 Conclusion

By filtering spammer reviews, we prepare a clean dataset. Based on that, we build AI tools to help businesses save time in replying to consumers' reviews and getting useful information and insights from reviews. In addition, the tools provide consumers convenience in extracting key information from reviews. With all the above, we achieve the goal of empowering the B to C platforms on efficiency, accuracy, retention, and healthy development.

References

[1] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. Text summarization techniques: A brief survey. *arXiv preprint arXiv:1707.02268*, 2017.

[2] Bruno Blanchet, Vincent Cheval, and Véronique Cortier. Integrating automated and interactive protocol verification. *arXiv preprint arXiv:2008.08612*, 2020.

[3] Ahmed M. Elmogy, Usman Tariq, Ammar Mohammed, and Atef Ibrahim. Fake reviews detection using supervised machine learning. *International Journal of Advanced Computer Science and Applications*, 2021.

[4] Sherry He, Brett Hollenbeck, and Davide Proserpio. The market for fake reviews. *Marketing Science*, 41(5):896-921, 2022.

[5] D. Klein and C.D. Manning. Accurate unlexicalized parsing. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 2003.

[6] Naveen Kumar, Deepak Venugopal, Liangfei Qiu, and Subodha Kumar. Detecting review manipulation on online platforms with hierarchical supervised learning. *Journal of Management Information Systems*, 35(1):350-380, 2018.

[7] M. Ott, Y. Choi, C. Cardie, and J.T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

[8] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*, 2011.

[9] Yubao Zhang, Shuai Hao, and Haining Wang. Detecting incentivized review groups with co-review graph. *High-Confidence Computing*, 1(1):100006, 2021.