Suitability Analysis of Text Analysis Tools for Food Service Industry Customer Reviews

700021737

I. Introduction

The food service industry relies heavily on customer reviews to attract new customers, as well as understand where improvements can be made in their business. Increasingly, social media can have a huge impact on a restaurant's financial success or failure [1]. Text analysis tools can provide huge insights towards understanding the opinions of customers, as well as the issues that they raise across large textual datasets. The aim of this project is to compare the suitability of text analysis tools in the fields of sentiment analysis and topic detection, applied to datasets of restaurant reviews. The project will aim to identify which sentiment analysis tool is best for predicting the sentiment of a user review compared to their declared star rating, across an international textual dataset. It will also evaluate tools for topic detection, with the goal of allowing industry decision makers access to information about the root causes of users positive or negative reviews.

Previous work has been performed on review sentiment analysis, focusing on how a sentiment model handles reviews from different cultures [2]. They found that the model experienced variance across cultures, due to the differing ways that opinions are expressed. This project will build from this finding, by determining whether different models are better suited to different cultures reviews.

II. Datasets

The datasets used in this project consist of restaurant reviews written by the general public and posted online. Each review is labelled with a self declared star rating between 1 and 5.

A. Tripadvisor Scraped Dataset

For international comparison, this project required a dataset of food service reviews labelled with country of the establishment. A suitable dataset was not available, therefore a custom collection process was undertaken. The Scrapy [1] Python library was used to collect reviews from Tripadvisor [2], using an automated requests framework and HTML parser based on CSS tagging. The end result was a reasonable sized dataset in .JSONL format. The distribution of reviews can be seen in table I.

B. Kaggle Dataset

While the Tripadvisor data collection was successful, it was very time consuming, especially to collect enough low starred reviews. Therefore, for some of the later experiments, a publically available dataset of 10,000 Indian resturants from Kaggle is used [3]. It should be noted that this dataset contains half star ratings, in preprocessing these are rounded up to the nearest whole star (Table II).

Country	Total Reviews	5	4	3	2	1
USA	792	44%	31%	13%	6%	6%
UK	803	56%	17%	6%	6%	15%

TABLE I
Star rating distribution across country of origin

Total Reviews	5	4	3	2	1
9954	39%	25%	12%	7%	17%

TABLE II
Star rating distribution

III. Sentiment Analysis

Three sentiment analysis tools were used to evaluate their usefulness in prediction of star rating using sentiment: VADER [4], TextBlob [3], and BERT [5].

A. Sentiment Calculation

The reviews for each dataset were passed to each of the sentiment analysis tools. Each review was also put into a category "Good" = 2 "Neutral" = 1 and "Bad" = 0, due to the tools finding it difficult to distinguish between close ratings such as 4/5 or 1/2 stars. Figure 1 shows two of the review instances after sentiment analysis has been applied.

review	stars	category	vader	blobpolar	blobsubj	bert_label	bert_score
Good food, prompt service, Murtaza was particularly helpful and made me feel special, food was delicious	5	2	0.908	0.555952	0.62619	1	0.999882
Went to kfc sidwell Street to treat my grandson. Wait time wasn't bad. Unfortunately that was the best part of the visit. Chicken was dry and barely warm. Ended up leaving...	1	0	0.8934	-0.00833	0.63333	0	0.997065

Fig. 1. Example reviews with sentiment scores

[1] https://scrapy.org/
[2] https://www.tripadvisor.co.uk/
[3] https://textblob.readthedocs.io/en/dev/

B. Predicting Category Using Sentiment

A machine learning decision tree model was fitted to the dataset for each tool, using the sentiment scores from each tool as features, and the category as classification target. The accuracy scores for each model and dataset are shown in Figure 2. From this we can see that TextBlob was the most accurate predicting sentiment on the UK reviews, wheras BERT was the most accurate at predicting the USA reviews. This experiment has started to show interesting results, however it should be noted that due to the relatively small sample size, this may be due to noise in the data. Further study with more resources for data collection is likely to provide more reliable results. Furthermore, the imbalance of review star distributions (Table I) may have lead to the misclassification of the minority classes of negative and neutral reviews, seen in the confusion matrices (Figures 3, 4).

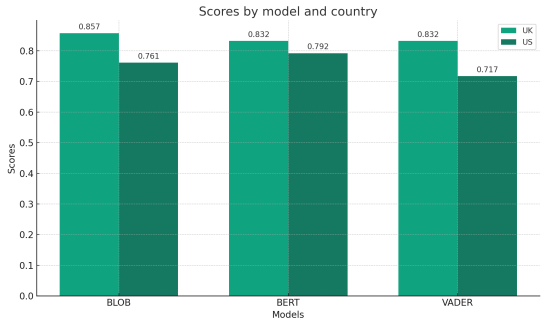

Fig. 2. Model Accuracy Scores for each sentiment algorithm and country

Fig. 3. Confusion Matrix for TextBlob using UK reviews

IV. Lexical analysis

Lexical analysis can provide useful insights into the ways that users express themselves in reviews, and can be used to aid sentiment analysis. Breaking text down into its basic parts can allow identification of sentiment expressive words, and judge the intensity of the expression through the type of adjectives used for example. From this point forward in the project, the Kaggle Indian resturants dataset [3] was used due to its larger size.

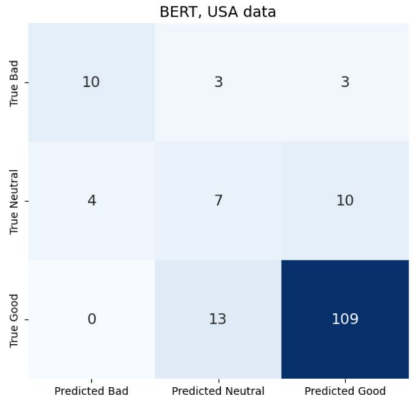

Fig. 4. Confusion Matrix for BERT using USA reviews

A. Adjective extraction

Part-of-Speech tagging was used to extract adjectives from the dataset, to discover what were the common adjectives used in both the positive and negative reviews. The text was first pre-processed, including lowercasing, removing stopwords and punctuation, then tokenising. Adjective extraction was performed using the Natural Language ToolKit. Table III shows the most commonly used adjectives in positive (4/5 star) reviews, and table IV shows the adjectives from negative reviews. The negative reviews highlight a common pitfall for this type of analysis, whereby ”good” is found to be the most common adjective even for negative reviews, because it is frequently used with a negation ie ”not good”. This can potentially be avoided by using bigram analysis.

Adjective	Frequency
”good”	5090
”great”	1601
”nice”	1140
”best”	1086
”awesome”	608

TABLE III
MOST POPULAR ADJECTIVES IN POSITIVE REVIEWS

Adjective	Frequency
”good”	733
”bad”	520
”worst”	431
”pathetic”	176
”disappointed”	112

TABLE IV
MOST POPULAR ADJECTIVES IN NEGATIVE REVIEWS

V. Topic Detection

Being able to analyse the topics discussed in online reviews can provide key insights to a number of groups. Firstly, industry managers can monitor the topics discussed in reviews of their business, and better allocate efforts to areas that need the most improvement. Another area of possibility is in food standards enforcement, a government agency could use an automated system that monitors reviews for restaurants in their area. If a restaurant has a sudden spike in reviews on the

topic of poor hygiene, they can be scheduled for a surprise inspection. This project focused on five common categories found in both positive and negative reviews: food quality, service quality, hygiene, restaurant ambiance and price. The techniques are aimed at detecting whether or not the review mentions each of the topics.

A. Keyword detection

One of the simplest methods for topic detection is to use keywords. This project experimented using a set of pre-defined keywords for each topic, and would return true for that category if the keyword was present.

B. GPT

A far more advanced approach was performed using the GPT-3.5 Turbo model [6]. The reviews were fed into the OpenAI API, and the model was asked to return which categories were present in the review. Only a subset of 2000 of the 10000 total reviews were labelled by GPT due to time and expense constraints.

C. Comparison between keyword and GPT results

Tables V and VI show examples of the differing opinions between the keyword detector and GPT 3.5. Figure 5 shows the recall for each catagory, and highlights how GPT is able to identify far more of the reviews as containing the topic.

"Ordered food via swiggy and got old and rubbery chicken. Of course it was stale and leftover stuff which was remodeled and resold. Shameful conduct, never ordering again."					
Model	**Food**	**Service**	**Hygiene**	**Ambiance**	**Price**
Keywords	True				
GPT 3.5	True		True		

TABLE V
EXAMPLE LABELLED REVIEW

"Actually I struggled between 3 and 4 for this place, where drinks are so good and buffet doesn't last to it. But a totally classy atmosphere in buffet area and very good vibes in first floor and their brewery is one of the best. A little costly. Caution: if you are a stag majority of times they won't allow u on ground floor in evenings, which is kind of sucks though. Finally value for money for a date or with a team else little expensive"					
Model	**Food**	**Service**	**Hygiene**	**Ambiance**	**Price**
Keywords					True
GPT 3.5	True			True	True

TABLE VI
EXAMPLE LABELLED REVIEW

D. Topic insights from GPT

Now that the reviews have been labelled with their estimated contents, some interesting experiments can be conducted. Figure 6 shows the percentage of negative and positive reviews that contain each class. From this we can see that negative reviews are far more likely to discuss hygiene, whereas positive reviews are more likely to discuss the ambience of the restaurant. This type of analysis can give an insight into which aspects of the dining experience customers hold as most important.

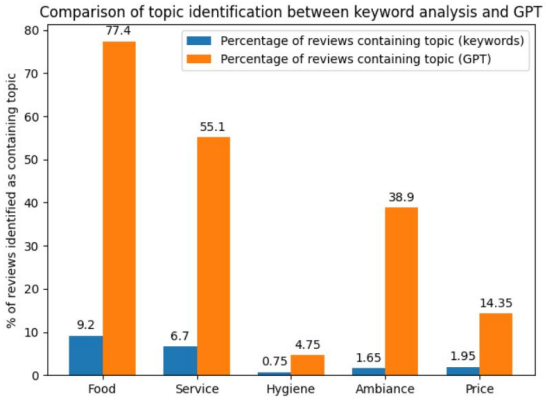

Fig. 5. Percentage of reviews identified to contain each topic by keyword analysis and GPT

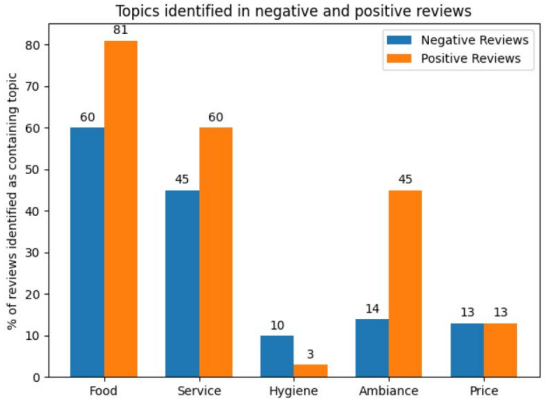

Fig. 6. Percentage of negative and positive reviews determined to contain each topic

VI. Conclusion

Online reviews are vital for any business, especially in the resturant industry. For large businesses looking to gain an insight into their own and their competitors customer opinions through online reviews, text analysis tools can provide quick and insightful information. This project has found that different sentiment analysis tools are better suited to predictions of sentiment across cultures. Initial results suggest that TextBlob works best on data from the UK, and BERT performs best on US data, however larger datasets are needed to confirm this.

Lexical analysis was also investigated, highlighting some of the pitfalls when dealing with subjective text such as reviews and how people choose to express their opinions. Finally, topic detection was used to investigate which parts of the dining experience the user commented on. Powerful models such as GPT-3.5 have the ability to detect very nuanced mentions of topics, and this was used in this project to determine topics that were more present in positive or negative reviews. It was found that negative reviews are far more likely to discuss hygiene issues, wheras positive reviews are more likely to comment on ambience.

As the field of text analysis progresses, it will be vital for business leaders to keep up with the latest tools to ensure a competitive edge, and apply the correct tools for the task.

REFERENCES

[1] S. M. Fernández-Miguélez, M. Díaz-Puche, J. A. Campos-Soria, and F. Galán-Valdivieso, "The impact of social media on restaurant corporations' financial performance," *Sustainability*, vol. 12, no. 4, 2020. [Online]. Available: https://www.mdpi.com/2071-1050/12/4/1646

[2] Y. Wan and M. Nakayama, "A sentiment analysis of star-rating: a cross-cultural perspective," 01 2022. [Online]. Available: https://www.researchgate.net/publication/357748212

[3] J. ARVIDSSON, "10000 Restaurant Reviews," https://www.kaggle.com/datasets/joebeachcapital/restaurant-reviews, 2023, [Accessed: 10/03/2024].

[4] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1. AAAI, 2014, pp. 216–225. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14550

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," in *Advances in neural information processing systems*, vol. 33, 2020, pp. 1877–1901.