Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

A Cantonese Restaurant Review Dataset for Aspect Category Sentiment Analysis with XLM-RoBERTa

Yuxin Cai *

Chinese International School, Hong Kong, China

* Corresponding Author Email: cicic2025@student.cis.edu.hk

Abstract. The volume of online consumer reviews has surged as e-commerce continues to grow, providing valuable insights for both consumers and business owners. Aspect Category Based Sentiment Analysis (ACSA) identifies sentiment polarity based on specific aspect categories in reviews. Despite extensive research in other languages, Cantonese remains underexplored due to its unique linguistic features and limited datasets. This study seeks to bridge this gap by fine-tuning a RoBERTa-based model for Aspect Category Sentiment Analysis (ACSA) on Cantonese restaurant reviews from OpenRice. A dataset is constructed by collecting and annotating 7,473 reviews based on five aspect categories: food, service, ambience, price, and timeliness. The fine-tuned XLM-RoBERTa model achieved an accuracy of 75.17%, outperforming five baseline models and demonstrating the efficacy of transformer-based models in low-resource languages. The study shows that the fine-tuned RoBERTa-based model has significant advantages in processing low-resource languages such as Cantonese, not only surpassing the baseline model in accuracy, but also providing a solid foundation for future research on Cantonese sentiment analysis. This work contributes a significant dataset and highlights potential future research directions.

Keywords: Cantonese natural language processing; sentiment analysis; multilingual Roberta.

1. Introduction

As e-commerce became more popular, there is an increase in the quantity and availability of online consumer reviews [1]. It offers valuable insights into the user sentiments towards products to consumers and business owners. Nevertheless, comprehending the sentiments expressed across diverse customer feedback can be arduous. Aspect-based sentiment analysis (ABSA) can address this challenge by systematically identifying the sentiment polarity of a given user reviews based on specific aspect terms in the text. Aspect category-based sentiment analysis (ACSA) is a type of ABSA that focuses on the sentiment of an aspect category. For instance, in the review "The food is great, but the service is terrible", the sentiment of the first aspect category "food" is positive and the second aspect category "service" is negative.

Cantonese, a Sinitic language spoken by about 70 million people [2]. It is significantly less explored for sentiment analysis compared with other languages such as English and Chinese [3]. This is partly due to the scarcity of available datasets, as well as Cantonese's unique linguistic features, including the language's lexical complexity, with many polysemous words and idioms, and its grammatical differences from IndoEuropean languages, such as the use of sentence particles and classifier system [4].

In Cantonese natural language processing tasks, earlier works focused on more traditional supervised machine learning approaches. For example, Zhang et al. explored Support Vector Machine (SVM) and Naive Bayes classifiers for sentiment classification in 1,800 online Cantonese reviews [5]. However, due to the limited dataset of their study, the model may lack generalisability. More recently, studies focus on lexicon-based methods for Cantonese sentiment analysis. Klyueva et al. specifically targeted food-related sentiment, obtaining a Cantonese sentiment lexicon containing 1,887 positive words and 858 negative words [6]. Xiang et al. constructed a Cantonese sentiment lexicon using an automatic method, and incorporated lexical knowledge into the attention layer of a Long-Short Term Memory (LSTM) model, achieving an accuracy of 60.8% [7]. Despite various attempts to enhance

Cantonese sentiment analysis, previous methods generally achieve suboptimal performance, and Cantonese ACSA remains relatively underexplored.

Considering the shortage of language resources for Cantonese ACSA and the pressing need for sentiment analysis among businesses in Hong Kong and the Guangdong region, this study fine-tunes a RoBERTabased model to perform ACSA. Specifically, this work examines restaurant reviews on OpenRice, a popular dining guide platform with 7.5 million active users and 500,000 restaurants registered [8]. 7,473 randomly selected Cantonese reviews are scraped from OpenRice with their aspect category-based sentiment polarity annotated. To the best of current knowledge, this dataset is the largest Cantonese review dataset for ACSA tasks. The reviews were analysed using a fine-tuned multilingual RoBERTa model, given its exceptional capability to handle complex linguistic features [9]. Results suggest the multilingual RoBERTa (XLM-RoBERTa) model achieves high performance across all aspect categories.

2. OpenRice ACSA Dataset

2.1. Dataset Construction

Due to the lack of an existing large-scale Cantonese review dataset for ACSA tasks, this study contributes by constructing a specifically tailored Cantonese review dataset, which was named as the OpenRice ACSA Dataset. A total of 7374 reviews were scraped using Python web scraping tools from 225 randomly selected Hong Kong restaurants on OpenRice. Each review contains textual content and an overall rating of either “smile”, “ok”, or “cry”, which was treated as the overall sentiment label of the review, mapping to positive, neutral, and negative sentiments respectively. The reviews are primarily written in Cantonese (traditional Chinese characters), although some are mixed with English and Mandarin (simplified Chinese characters).

2.2. Aspect Categories

Table 1. Definitions of the 5 aspect categories.

Aspect Category	Food	Service	Ambience	Price	Timeliness
Definition	The quality, taste, Presentation, and variety of the food offerings.	The attentiveness, responsiveness, and professionalism of the staff.	The overall atmosphere, and environment of the establishment.	The Affordability and value of the menu items and dining experience.	The promptness and efficiency of the service and food preparation.

Since the scraped review only has an overall sentiment label regarding the restaurant, aspect categories are determined for the ACSA tasks. The 5 aspect categories annotated for this dataset are food, service, ambiance, price, and timelines. For example, in one of the reviews from the OpenRice ACSA Dataset, “It’s expensive and not delicious. The souffle takes a long time to come”, the aspect categories mentioned include food, price, and timeliness. Table 1 presents a full list of categories and their respective definitions.

2.3. Data Annotation

Since the scraped review only has an overall sentiment label regarding the restaurant, the sentiment expressed in each restaurant review towards specific aspect categories is annotated as 1 (Positive), 0 (Neutral), or -1 (Negative). It’s important to note that if a review does not explicitly mention an aspect category, a sentiment score of 0 (Neutral) is assigned by default for that category in the review.

First, the raw dataset is labelled using Llama3-8B-Chinese-Chat, a large language model fine-tuned by Wang et al. on a mixed Chinese-English dataset [10]. The prompt directs the llama to evaluate the sentiment (-1, 0, or 1) based on the 5 categories and provide a brief justification for each aspect category’s rating. Second, to ensure accuracy, the dataset labelled by Llama3-8B-Chinese-Chat is

randomly split into three subsets, and each subset is manually verified by a native Cantonese speaker given both the original review and Llama3-8B-Chinese-Chat's reasoning for each annotation. Third, each reviewer verifies the annotations of a different subset. Controversial restaurant reviews were discussed further until an agreement was reached.

2.4. Dataset Analysis

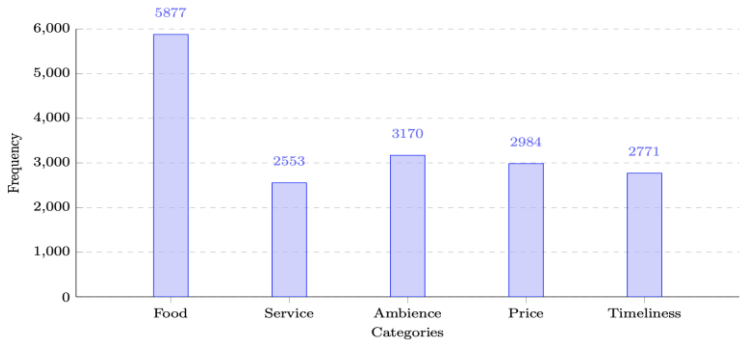

Fig 1. Distribution of categories (Photo credit: Original).

The dataset is completely balanced in terms of the reviews' overall sentiments, with each class (positive, neutral, and negative) having some samples of 2491. The longest restaurant review in the dataset has 666 characters, while the shortest review has 34 characters. In terms of the distribution of aspect categories, Fig. 1 shows that the food aspect category is the most prevalent, being mentioned in 78.6% of all the reviews, while service is mentioned only in 34.16% of the reviews. This is expected, as food is often the primary focus and reason for choosing a particular restaurant. Apart from the food category, the distributions for the other 4 aspect categories are relatively balanced. This particular characteristic of the dataset proves its utility in sentiment analysis related tasks.

3. Methodology

3.1. Problem Formulation

The ACSA task is framed as a multi-label classification problem. Given a training dataset, D, for each input review sentence, denoted by x and represented by a sequence of tokens, $(t_1, t_2, \ldots, t_L)$, where L is the length of the sentence, is analysed to assess sentiments related to multiple predefined aspects. Denote A = $(a_1, a_2, \ldots)$ an as the set of aspect categories, where N is the total number of aspect categories and n is fixed at 5. Each aspect ai belongs to a specific category such as food quality, service, ambience, price, and timeliness. The goal is to predict the sentiment polarity $y_i \in \{-1, 0, 1\}$ for each review xi in x with respect to each aspect category a_i, Here, −1, 0, and 1 represent negative, neutral, and positive sentiments, respectively.

3.2. Model Architecture

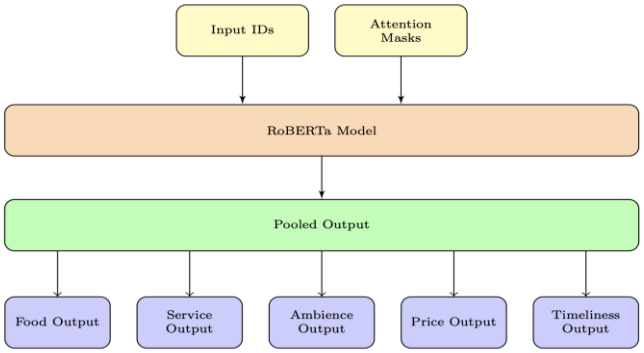

Fig 2. The architecture of the proposed model used for Cantonese ACSA (Photo credit: Original).

The architecture of the model used for review-based ACSA uses a Transformer-based approach, specifically leveraging the XLM-RoBERTa model. As shown in Fig. 2, the model incorporates several components that process inputs, extract features through the Transformer, and classify these features into distinct sentiment categories.

The model begins with two input layers designed to handle different aspects of the data: Input IDs and Attention Masks. At the heart of the model lies the XLM-RoBERTa transformer. This transformer processes the input IDs and attention masks through multiple layers of self-attention and feed-forward networks. The output from XLM-RoBERTa can be described by the following transformation within each self-attention layer:

$$\text{Attention}\ (\mathrm{Q}, \mathrm{K}, \mathrm{V}) \ = \ \text{softmax}(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}})\mathrm{V} \tag{1}$$

Q, K, and V represent the queries, keys, and values matrices, respectively, and $\mathrm{d_k}$ is the dimensionality of the keys. This mechanism enables the model to weigh the importance of different words irrespective of their position in the input sequence.

The output from XLM-RoBERTa includes a sequence of hidden states for each token and a pooled output. This pooled output is then utilized for classification tasks. The model branches into multiple dense layers, each corresponding to a different aspect of sentiment analysis (food, service, ambience, price, and timeliness). Each of these layers employs a softmax activation function, which can be represented as:

$$\text{softmax}(\mathrm{z_i}) \ = \ \frac{\mathrm{e}^{\mathrm{z_i}}}{\sum_{\mathrm{j}} \mathrm{e}^{\mathrm{z_j}}} \tag{2}$$

$\mathrm{z_i}$is the logit or raw score for each class. The softmax function converts these logits into probabilities by exponentiating and normalizing them, thus allowing for a probabilistic interpretation of the output across multiple classes.

3.3. Loss Function

The proposed model adopts a categorical cross-entropy loss function, quantifying the difference between the predicted and actual sentiment classifications. Formally, the loss for a single instance when predicting across multiple classes is defined as:

$$\mathcal{L}(\mathrm{y}, \hat{\mathrm{y}}) \ = \ -\sum_{\mathrm{c}=1}^{\mathrm{C}} \mathrm{y_c} \log(\widehat{\mathrm{y_c}}) \tag{3}$$

C is the number of classes (negative, neutral, and positive),

$\mathrm{y_c}$ is a binary indicator (0 or 1) if class label c is the correct classification for the observation,

$\hat{\mathrm{y}}$c is the predicted probability of the observation belonging to class C.

In the context of the ACSA task, this loss is computed separately for each aspecta_i, with each aspect treated as an independent multi-class classification problem. Therefore, the total loss for a given restaurant review is the sum of the losses for each aspect category:

$$\mathcal{L}_{\text{total}} \ = \mathcal{L}_{\text{food}} + \mathcal{L}_{\text{service}} + \mathcal{L}_{\text{ambience}} + \mathcal{L}_{\text{price}} + \mathcal{L}_{\text{timeliness}} \tag{4}$$

4. Experiment and Results

4.1. Data Preparation

The dataset, consisting of 7,473 Cantonese restaurant reviews from OpenRice as mentioned in section 2, undergoes a comprehensive preprocessing pipeline to enhance data quality for subsequent analysis. URLs are first removed to eliminate irrelevant web links. The text is then normalised using Unicode standards. Emojis are replaced with their descriptive text, enriching the linguistic content. Since some reviews contain English texts, all reviews are converted to lowercase, and punctuation is stripped to focus solely on words. Finally, stopwords are removed to exclude common but uninformative words from the analysis. The cleaned text is then split into training and testing subsets, with 20% of the data reserved for testing to validate the model's performance.

4.2. Implementation Details

The text data preparation involves using the XLMRobertaTokenizer [11], which supports the multilingual RoBERTa Model. This tokenizer converts each text in the training set into token indices and creates attention masks, standardising all sequences to a maximum length of 200.

The XLM-RoBERTa model is fine-tuned with specific configurations, using an Adam optimizer with a learning rate of 1e-5 and trained over 5 epochs with a batch size of 32, to gently adjust the pre-trained weights. The training process includes a 10% validation split to monitor generalisation and avoid overfitting, ensuring the model's robustness across diverse linguistic inputs, particularly tailored for the nuances of Cantonese text in a restaurant context.

4.3. Baseline Models

The To better evaluate the performance of the proposed model, 5 baseline models are implemented across different machine learning techniques and complexity levels. These models include Naive Bayes Classifier (NB), Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree, and Bidirectional Encoder Representations from Transformers (BERT) [12].

NB: This model applies Bayes' Theorem, assuming strong independence between the features. The multinomial variant of the Naive Bayes model is utilised for ACSA.

LR: Logistic regression estimates the probabilities using a logistic function. Each aspect category is treated as a separate classification problem where the model uses TF-IDF transformed text data for prediction.

SVM: a robust classifier that finds the optimal hyperplane which maximises the margin between different classes in the feature space. It is adept at handling non-linear relationships using kernel functions.

Decision Tree: Decision trees are a non-parametric method by learning decision rules from features. During training, each category is fitted by an independent Decision Tree Classifier.

BERT: BERT is a transformer-based machine learning technique for natural language processing pre-training. This study uses a multilingual BERT model that is pre-trained on a large corpus of multilingual data from Wikipedia.

4.4. Evaluation Metrics

In assessing the performance of models for Aspect Category-based Sentiment Analysis (ACSA), it utilises a range of evaluation metrics tailored to each aspect category, including weighted F1 score, weighted precision, weighted recall, and accuracy. Weight F1 score, weighted precision, and weighted recall are calculated as the average of the scores across the 5 aspect categories. Similarly, the accuracy of the model's performance is calculated by averaging the accuracy scores obtained for

each aspect category. This aggregated metric provides a holistic view of the model's effectiveness across all categories.

4.5. Main Results

Table 2. Model performances for Cantonese ACSA; best performances are highlighted in bold.

Model	Weighted			Acc. (categorical)					Acc.
	F1	Pre.	Rec.	Food	Service	Ambience	Price	Timeliness	
SVM	56.38	62.82	53.90	65.42	65.35	60.2	61.07	62.07	62.82
NB	47.47	62.09	48.06	62.07	64.82	59.00	61.07	63.48	62.09
Decision Tree	50.43	54.30	51.69	57.19	55.32	54.18	50.37	54.45	54.30
LR	57.24	62.10	53.23	64.28	65.69	59.06	59.87	61.61	62.10
BERT	69.59	71.28	66.32	71.97	74.78	75.05	68.29	66.29	71.28
XLM-RoBERTa	**74.58**	**75.13**	**73.79**	**76.72**	**79.80**	**77.26**	**73.04**	**69.03**	**75.17**

In the evaluation of various models for the Cantonese ACSA task, the fine-tuned XLM-RoBERTa model demonstrates superior performance across all metrics compared to other models such as SVM, NB, Decision Tree, LR, and BERT. The results, as summarized in Table 2, clearly indicate the effectiveness of the XLMRoBERTa model in handling this sentiment analysis task.

The RoBERTa model achieved the highest scores in all measured metrics, including Weighted F1, Precision, and Recall, with scores of 74.58, 75.13, and 73.79 respectively. XLM-RoBERTa scored 76.72 in "Food", 79.80 in "Service", 77.26 in "Ambience", 73.04 in "Price", and 69.03 in 'Timeliness'. This indicates the model's robust ability to generalize across different aspects of sentiment analysis. The superior performance of XLM-RoBERTa could be attributed to its architecture, which utilizes a robustly optimized BERT pre-training approach. Unlike BERT, RoBERTa-based models aretrained with a larger dataset and longer training times, and they modifies key hyperparameters in BERT, including removing the next-sentence pretraining objective and training with much larger mini-batches and learning rates. This might explain the enhanced ability of RoBERTa to understand and analyze the nuances of sentiment expressed in Cantonese, a context where linguistic subtleties are paramount.

Despite its overall superior performance, the XLM-RoBERTa model exhibits relatively lower performance in the "Price" and "Timeliness" categories with F1 scores of 73.04 and 69.03, respectively, compared to 76.72 and 79.80 in "Food" and "Service". This discrepancy may be due to the lesser amount of training data available for these categories, or the inherent complexity and variability of sentiments associated with pricing and service timeliness, as sentiments related to price and timeliness are often influenced by individual expectations and subjective value assessments, which can be more challenging to model accurately.

5. Analysis and Discussion

5.1. Ablation Study

Table 3. Performance metrics for different max_len; best performances are highlighted in bold.

max_len	Weighted			Acc.
	F1	Pre.	Rec.	
64	70.06	67.84	70.06	62.82
128	73.57	72.54	73.57	62.09
200	74.58	75.13	73.79	54.30
256	**76.88**	**76.43**	**76.88**	**62.10**

Considering the idea that the maximum length of the input tokens (max_len) can affect the performance of the XLM-RoBERTa model, an ablation study is conducted to find the optimal max_len. Table 3 illustrates the results when max_len is set to 64, 128, 200, and 256.

The results indicate that increasing the maximum token length generally improves model performance. Specifically, when max_len is set to 64, the model achieves a weighted recall, precision, and F1 score of 70.06, 67.84, and 70.06 respectively, with an accuracy of 66.41%. As the token length increases to 128, there is a noticeable improvement in the metrics, with the F1 score reaching 73.57 and accuracy reaching 72.11%. The best performance is achieved with a maximum token length of 256, yielding a weighted recall, precision, and F1 score of 76.88, and an accuracy of 76.24%. These results highlight the importance of selecting an appropriate max_len for optimizing model performance. By setting the max_len to 256, the fine-tuned XLM-RoBERTa model achieves its highest performance across all metrics.

5.2. Discussion

Overall, XLM-RoBERTa's sophisticated architecture can effectively manage the intricacies of Cantonese in the context of restaurant review-related ACSA. The model's ability to leverage larger and more diverse training datasets, along with longer training periods, enables it to grasp nuanced linguistic features that are specific to Cantonese. The robust performance of the fine-tuned model also hints that the constructed ACSA dataset provides a comprehensive representation of the various sentiment aspects in Cantonese restaurant reviews, contributing to the model's effectiveness.

Despite its strengths, the proposed method has limitations for Cantonese ACSA. As the maximum token length increases, the computational complexity and resource requirements also increase, making it challenging to process very long texts efficiently. This can lead to issues with memory usage and processing time, potentially limiting the model's applicability in real-time or resource-constrained environments. Furthermore, since the unmentioned aspect categories are labeled as 0, which is the same label used for neutral sentiment, this can cause ambiguity in the classification results. One solution is to apply a mask that indicates the presence of the aspect category, thereby differentiating between truly neutral sentiments and unmentioned aspects.

6. Conclusion

In conclusion, this paper presents a comprehensive Cantonese review dataset containing 7473 reviews, each annotated for sentiment polarity based on five distinct aspect categories. This dataset is aimed at facilitating advancements in natural language processing within the context of the Cantonese language. By making this dataset publicly available, the author hopes to encourage further research and development in this area. This paper further explored the performance of several widely used machine learning models on this new dataset. Among these models, XLM-RoBERTa achieved the highest accuracy, with a notable performance of 75.17%. This indicates the potential of transformer-based models in handling sentiment analysis tasks in low-resource languages like Cantonese. For future work, one potential avenue is the integration of userrelated information, such as user profiles or historical review data, to provide additional context that might improve sentiment prediction accuracy. Another direction involves the use of semantic and syntactic masks, which could help the models better understand and capture the intricate nuances of the Cantonese language.

References

[1] Y Tian Y, Stewart C M. History of E-Commerce. ResearchGate, 2007.

[2] Ngai E W T, Lee M C M, Choi Y S, et al. Multiple-Domain Sentiment Classification for Cantonese Using a Combined Approach. AIS Electronic Library (AISeL), 2018.

[3] Lee J. toward a Parallel Corpus of Spoken Cantonese and Written Chinese. 2011.

[4] Alderete J, Chan Q, Chan M, et al. Cantonese grammar synopsis. 2017.

[5] Zhang Z, Ye Q, Zhang Z, et al. Sentiment classification of Internet restaurant reviews written in Cantonese. ResearchGate, 2011.

[6] Klyueva N, Long Y, Huang C R, et al. Food-related sentiment analysis for Cantonese. PolyU Scholars Hub, 2018.

[7] Xiang R, Jiao Y, Lu Q. Sentiment Augmented Attention Network for Cantonese Restaurant Review Analysis. PolyU Scholars Hub, 2019.

[8] OpenRice. Highlights / Key Facts. OpenRice, 2024.

[9] Liu Y. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv.org, 2019.

[10] Wang S, Zheng Y. Shenzhi-wang/Llama3-8B-Chinese-Chat · Hugging Face. Huggingface. co, 2024.

[11] Conneau A, et al. Unsupervised Cross-lingual Representation Learning at Scale. arXiv.org, 2019.

[12] Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv.org, 2018.