INFO 555 - Applied NLP Project

Kiwoon Hong

1 Description

Customer reviews play a crucial role in providing insights for most industries because they directly affect brand reputation, influence potential buyers, and inform strategic decision-making. Analyzing reviews helps in understanding detailed consumer feedback, allowing stakeholders to enhance product development, strengthen marketing strategies, and identify potential market trends.

This project focuses on analyzing car reviews using a combination of sentiment analysis and zero-shot classification techniques. The high-level goal for this project is to extract main consumer feedback from reviews. To achieve this, negative consumer feedback is extracted from the review data and categorized into several groups to identify complaint trends. This statistical approach interprets feedback patterns based on neural network-based emotion scores, making it easier to analyze customer sentiments in car reviews.

The Car Review Dataset[1] on Hugging Face contains a variety of car user reviews, including vehicle titles, ratings, and review texts. In this data, there is no category of the review that is the goal of this project, so the models used in this project are unsupervised learning.

2 Related Work

With the rapid development of AI, emotion analysis using transformer models has not been so long. paper by Guha et al. (2015), an aspect-level sentimental analysis using a classical statistical method was performed. A paper by Tran et al. (2022) analyzed various aspects of customer reviews were analyzed using neural network architectures before transformers such as RNN and CNN in e-commerce environments. However, paper by Shangipour ataei et al. (2020) also performed sentimental analysis based on BERT to extract the relationship between context dialogue sequence and response and determine whether the response is sarcastic.

3 Procedure

3.1 Experiment

Initially, data is extracted from the Hugging Face dataset, which contains a variety of consumer reviews about the care they purchased along with their corresponding ratings. The ratings are organized into discrete levels ranging from 1 to 5 while the reviews are consist of natural languages(English). For example:

```
"its not what you want its what you need!
Still the perfect car"
```

The number of ratings was not evenly distributed, with high ratings of 58.6% out of 5, 27.6% out of 4, 7.8% out of 3, 4.1% out of 2, and 2% out of 1. The majority of reviewers gave high ratings, which meant that the sentimental analysis was likely to be biased. To ensure balanced analysis, the project samples an equal number of reviews from each rating category (1, 2, 3, 4, and 5), totaling 3,725 reviews, with 745 samples per rating.

Next, the sentiment analysis is conducted using the Hugging Face sentiment-analysis pipeline. This step involves determining the positivity or negativity of each review and assigning a sentiment score. Only negatively labeled reviews were filtered because they were considered to be dissatisfied reviews. Reviews that yield a negative sentiment score exceeding a threshold of 0.7 are classified as consumer complaints. To compare model performance, in addition to the default `sentiment-analysis` pipeline, the `twitter-roberta-base-sentiment-latest`[2] model was also used in addition to the default pipeline. This model is a model for the text

[1] https://huggingface.co/datasets/florentgbelidji/car-reviews

[2] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

classification pipeline that produces neutral sentiment in addition to positive and negative.

To categorize these complaints, the project defines seven major categories: performance, efficiency, safety, comfort, design, price, and service. Utilizing the Hugging Face zero-shot-classification pipeline with the model `facebook/bart-large-mnli`[3], the project performs zero-shot classification to assign these categories to the extracted complaints.

Finally, the project calculates and ranks the average scores for each complaint category, focusing on the categories with the highest and second highest scores. The results will provide valuable insights into consumer dissatisfaction trends, enabling stakeholders to make informed decisions about product improvements and marketing strategies.

3.2 Evaluation & Analysis

3.2.1 Sentiment Analysis

The prediction scores of the sentiment analysis models were indirectly evaluated by comparing them to the ratings provided by customers in the review data, as there is no golden label. Also, a heuristic-based evaluation was performed as a complementary evaluation. Twenty review data were sampled per each model to directly determine the sentiment of the text. For indirect evaluation, the confusion matrices of sentiment analysis model 1 and model 2 were compared to analyze the tendency, accuracy, F1 score, etc. of the predictions. To do this, the scores need to be scaled to map a numeric sentiment score from 0 to 1 to a categorical rating of 1,2,3,4,5. So in model 1 with labels ’positive’ and ’negative’, the labels were mapped like this:

```
if label == 'NEGATIVE':
    rating = round(5 - score * 4)
else:
    rating = round(score * 4 + 1)
```

With this mapping, high ’negative’ scores are matched to a rating of 1, and high ’positive’ scores are matched to a rating of 5. Model 2 has an additional rule because it has an additional ’neutral’ label:

```
else:
    rating2 = round((score2 - 0.3)
    / 0.7 * 4 + 1)
```

The neutral label is different from the other labels, so that it maps a minimum value(0.35...) and maximum value(0.99...) to a rating of 1,2,3,4,5, based on the distribution of scores. Finally, a confusion matrix was created based on the predicted (mapped scores) and actual labels (ratings). Each confusion matrix can be found in the Appendix. For the heuristic-based evaluation, the researcher annotates a score of 1 if they believe they have appropriately predicted the predicted label, i.e. [’positive’, ’negative’] for model 1 and [’positive’, ’neutral’, ’negative’] for model 2, and 0 otherwise.

3.2.2 Zero-Shot Classification

Zero-shot classification model is evaluated by a heuristic-based method to randomly select a sample of 20 reviews for manual review. Each filtered negative review is assigned a score corresponding to seven complaint categories[performance, efficiency, safety, comfort, design, price, and service] by the Zero-shot classifier, which displays the two highest-scoring categories for each review. The researcher then annotates a score of 2 if two categories are correctly predicted, 1 if one, and 0 if none.

4 Results

Heuristic-based evaluations are attached in the appendix.

4.1 Sentiment Analysis

For Model 1, the confusion matrix (attached in the Appendix) shows a rather extreme tendency for negative reviews to map to a score of 1 and positive reviews to a score of 5. This is because the distribution of the Sentiment prediction score is [0.5, 1] rather than [0, 1]. Model 2 has fewer of these extremes, but at the expense of overall accuracy. From this, it can be concluded that the Score-Rating mapping method for evaluating Models 1, 2, and 3 is not very appropriate. The heuristic-based evaluation had a smaller sample but was more robust. Most of the models predicted sentiment correctly, with only two out of 20 failures for Model 1 and only one for Model 2.

4.2 Zero-Shot Classification

Zero-shot classification correctly predicted an average of 1.55 out of 2 categories in 20 samples. However, there were 3 out of 20 cases where sarcastic expressions like "`Avoid this car, unless`

[3] https://huggingface.co/facebook/bart-large-mnli

`you have money to throw away."` were misinterpreted as the 'price' category.

References

Satarupa Guha, Aditya Joshi, and Vasudeva Varma. 2015. Siel: Aspect based sentiment analysis in reviews. In *Proceedings of the 24th International Conference on World Wide Web*, pages 117–118. ACM.

Taha Shangipour ataei, Soroush Javdan, and Behrouz Minaei-Bidgoli. 2020. Applying transformers and aspect-based sentiment analysis approaches on sarcasm detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 67–71, Online. Association for Computational Linguistics.

Quang-Linh Tran, Phan Thanh Dat Le, and Trong-Hop Do. 2022. Aspect-based sentiment analysis for Vietnamese reviews about beauty product on E-commerce websites. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 767–776, Manila, Philippines. Association for Computational Linguistics.

A Appendix

Rating	Label	Score	Annotation
2	N/N	0.99/0.88	O/O
3	N/-	0.99/0.53	O/O
1	N/N	0.99/0.77	O/O
2	N/N	0.99/0.74	O/O
2	N/N	0.99/0.90	O/O
5	P/P	0.99/0.97	O/O
4	P/P	0.96/0.83	O/O
5	P/P	0.99/0.69	O/O
5	P/P	0.97/0.78	X/X
1	N/N	0.99/0.92	O/O
2	N/N	0.99/0.84	O/O
3	N/-	0.99/0.50	X/O
4	P/P	0.99/0.75	O/O
1	N/N	0.99/0.74	O/O
3	N/N	0.99/0.76	O/O
2	N/N	0.99/0.83	O/O
1	N/N	0.99/0.82	O/O
1	N/N	0.99/0.69	O/O
4	N/N	0.99/0.80	O/O
2	N/N	0.99/0.82	O/O

Table 1: Heuristic-based evaluation of sentiment analysis models 1 and 2. 'slash' is used to distinguish between models 1(left) and 2(right). Annotation is the researcher's annotation for the model's prediction.

	1	2	3	4	5
1	704 / 2	10 / 100	3 / 122	8 / 382	17 / 136
2	704 / 9	17 / 92	2 / 147	5 / 397	14 / 97
3	655 / 8	17 / 136	12 / 208	9 / 330	49 / 60
4	306 / 6	21 / 129	21 / 166	24 / 270	370 / 171
5	110 / 5	21 / 62	21 / 97	23 / 189	567 / 389

Table 2: Confusion Matrix: Sentiment analysis model 1 / 2 - Each row represents the predicted class, and each column represents the actual label.

Model	Accuracy	F1 score
Model 1	25.4%	0.144
Model 2	25.9%	0.231

Table 3: Accuracy & F1 Score for each sentiment analysis model: Score-Rating mapping method.

Category1	Score1	Category2	Score2	Annotation
design	0.98	service	0.52	1
perfor	0.91	price	0.77	2
safety	0.98	design	0.94	1
safety	0.93	perfor	0.27	2
safety	0.99	price	0.98	1
price	0.35	safety	0.29	1
perfor	0.68	design	0.54	1
service	0.97	perfor	0.57	2
service	0.85	perfor	0.55	2
safety	0.92	perfor	0.82	2
service	0.71	perfor	0.49	2
design	0.87	comfort	0.86	1
perfor	0.51	service	0.39	2
perfor	0.97	price	0.44	1
service	0.69	perfor	0.68	2
safety	0.76	perfor	0.66	2
price	0.84	perfor	0.81	1
service	0.93	price	0.92	2
price	0.94	perfor	0.85	1
perfor	0.86	safety	0.67	2

Table 4: Heuristic-based evaluation of Zero-shot classification model. 'perfor' corresponds to the 'performance' category. Annotation is the number of correct predictions made by the model, as judged by the researcher.