Natural language sentiment analysis and modeling on Walmart customer reviews

Wenhui Wang

Problem Statement:

The Walmart Customer Reviews Dataset offers a wealth of insights into consumer sentiment and product feedback related to one of the world's largest retail giants. This dataset contains a vast collection of customer reviews, star ratings, and other relevant information that has been gathered through web scraping and data compilation. The key features include: Customer Reviews: Detailed textual reviews provide firsthand accounts of shopping experiences and product satisfaction. Star Ratings: Each review is accompanied by a star rating, allowing for sentiment analysis and product rating assessment. Review Dates: The dataset includes review submission dates, facilitating temporal analysis and trend detection. Product Identification: For some reviews, product identification details such as SKU numbers or product categories are provided.

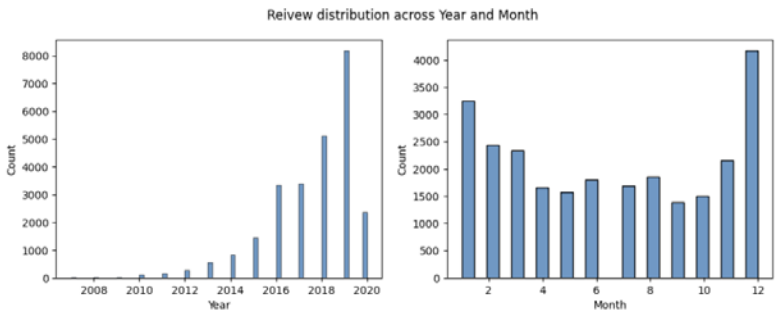
Based on this data set, I implemented sentiment analysis to find the key features that will promote or impair customers' experience. These information will not only help Walmart but also other retailers to improve their services. Secondly, I built model to predict customer rating based on review language by transfer learning and fine tuning on of pre-trained RoBERTa model (cardiffnlp/twitter-roberta-base-sentiment-latest) from hugging face.

Data Wrangling

I found two independently collected Walmart review data. There are only 1009 reviews overlapped between the two data sets. I decide to use one to build the model and use the other one as independent test. The two data sets both contain around 30k reviews (30006 vs 29997). I used the first one for data wrangling and exploration. After dropping 6 unrelated features, I got 13 features including Uniq id, Crawling time, pageurl, website, title, rating, review, reviewer name, review upvotes, review downvotes, verified purchaser, review date, etc . After remove records containing NA, I got 25822 records.

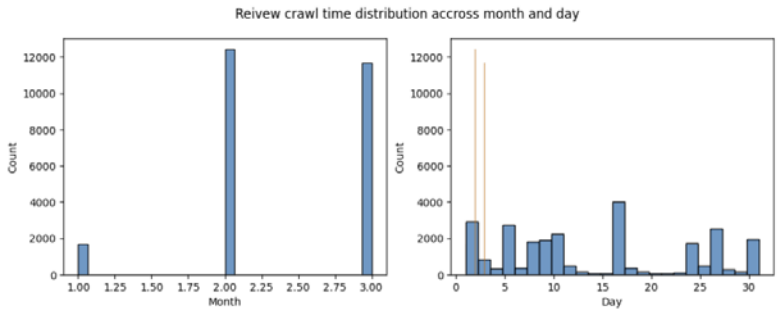
Exploratory Data analysis

I firstly checked the distribution of review dates.

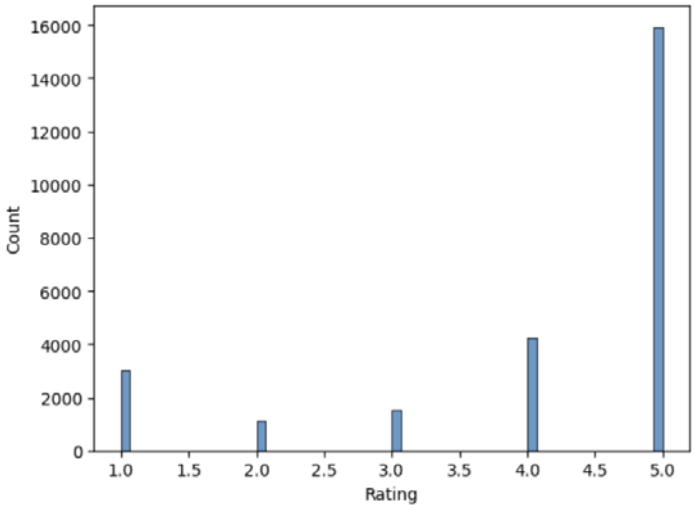


The figure indicate that the number of reviews increase with year. For year 2020, we only have reviews until March. The number of reviews in December and January is much higher than other months because these 2 months are holiday season.

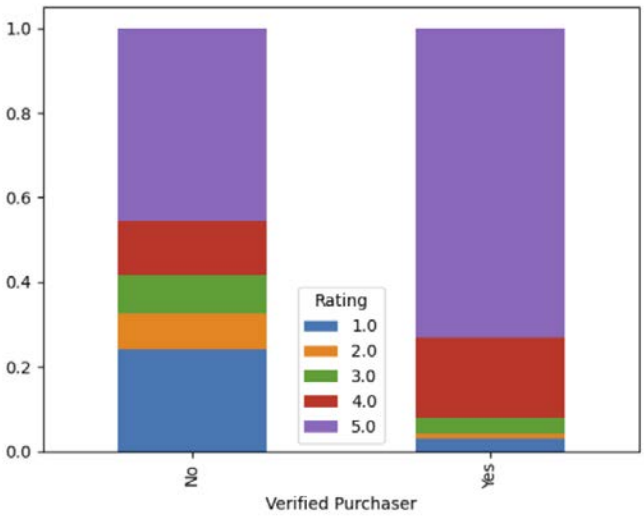
Secondly, I analyze the crawl time distribution. All the reviews are crawled in 2020. The time spread in the first 3 months.



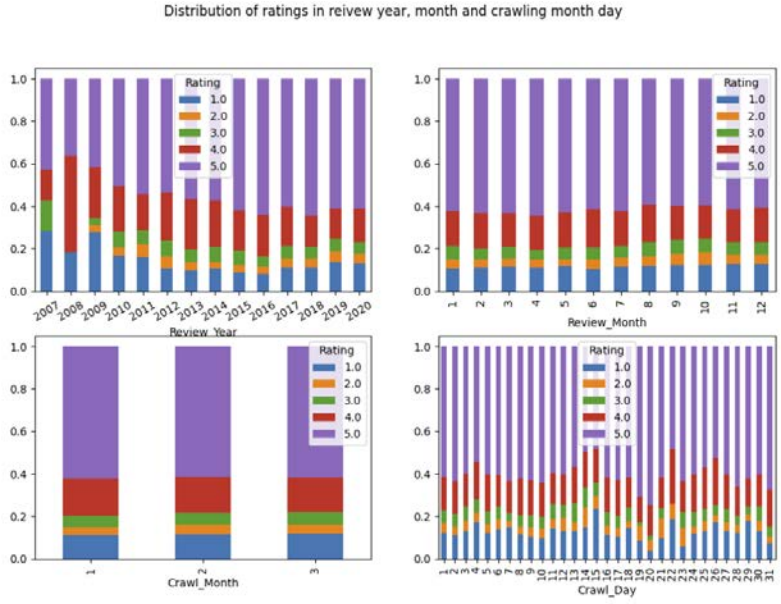
The ratings are ranging from 1 to 5. More positive ratings than negative ones.



Next, I checked the correlation between rating and other features. Verified purchaser give more positive rating than non-verified purchaser.

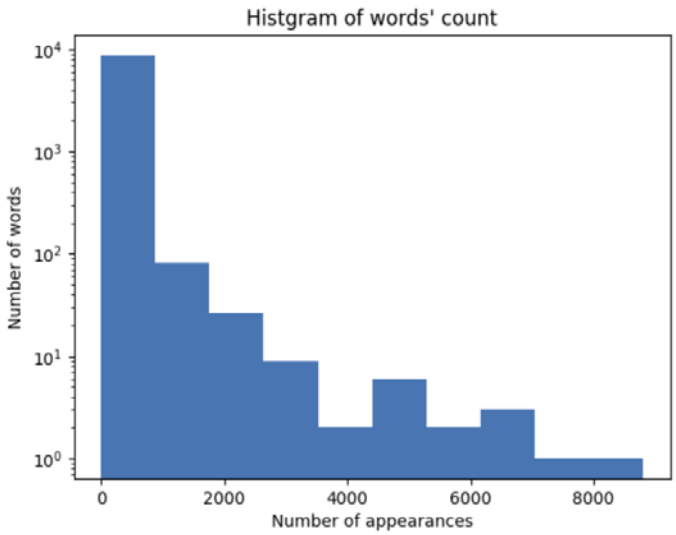


The ratings are not related with the review year/month and review crawling year/month.

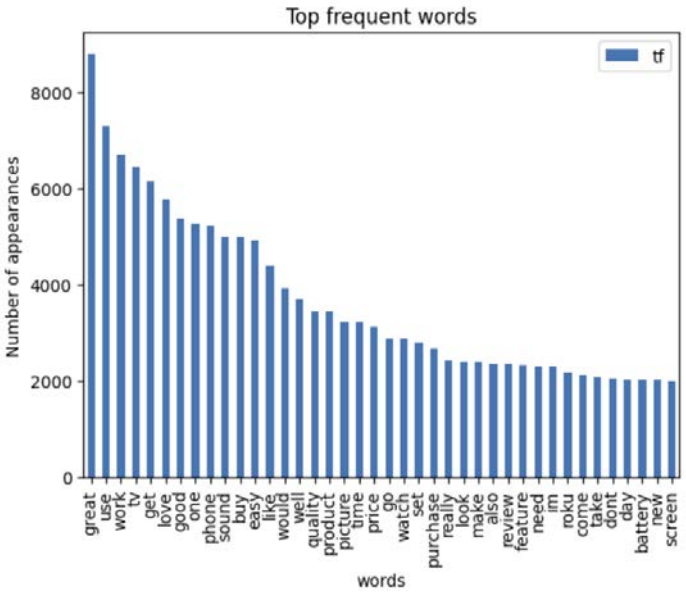


Sentiment analysis

At the beginning of sematic analysis, I ran text preprocessing which includes transforming to lower case, removing punctuations, number, stop words and rare words, lemmatization, etc. After the text preprocess, I ran exploration on distribution of words appearance.

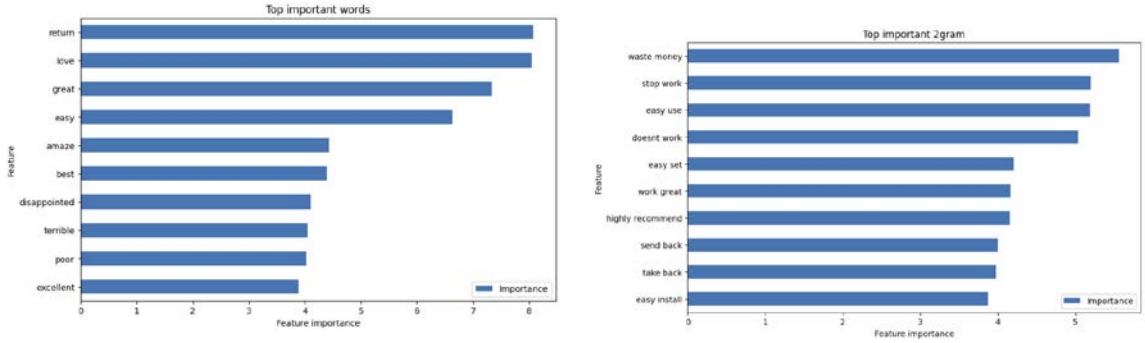


We can find that most of words appear less than 2000 times, only small number of words have high appearances. After that I checked the most frequent words (>2000). We can find that directly counting the number of appearances of words doesn't make much sense.



Word cloud which also base on number of appears indicate similar trend.

I dischotomized the ratings to positive (>3) and negative ratings (<=3). Then, I made use of logistic regression model to find the key features that will promote the rating or impair the rating. I split the data to train and test, after that I used tf_idf model trained by train data to encode train and test data sets. I checked both single word model and 2-gram model. The importance of a feather is defined as the absolute value from estimated parameters of the logistic model. The top ten features of single word model and 2-gram model are extracted.



Top import words and 2-grams indicate the most important words that indicate reviewer's rating and sentiment. These words align with common sense. At the same time, we can find that 2 grams are more informational on how to get positive review from customers.

Both single word and 2- gram logistic regression analysis indicate that the data contains information to model the association between reviews and ratings. 2-gram's performance is worse than single word may result from their much bigger feature space. But we can also find that 2-gram more much more informational than single word model for us to understand the key features of service and product that could result in positive ratings from customer. In such a case, we could expect that using more complex model such as BERT, we could have better modeling performance and at the same time, we could get more informational features to work on for improving service/product.

Model

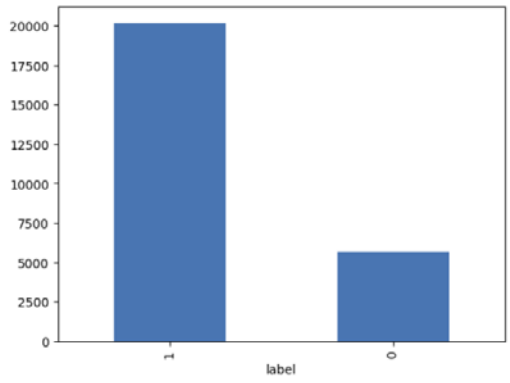
BERT[1], or Bidirectional Encoder Representations from Transformers , is a powerful natural language processing (NLP) model introduced by Google in 2018. It represents a breakthrough in pre-training techniques for language understanding tasks. BERT is built on the Transformer architecture, which is a type of neural network architecture that excels in capturing contextual information from sequential data. What sets BERT apart is its bidirectional approach to language understanding. Unlike previous models that processed language in a unidirectional manner, BERT looks at words in a sentence from both directions (left-to-right and right-to-left), allowing it to capture richer contextual information. This bidirectional pre-training helps BERT perform exceptionally well on a wide range of NLP tasks, including but not limited to, sentiment analysis, named entity recognition, question answering, and language translation.

RoBERTa[2], or Robustly optimized BERT approach, is a variant of the BERT (Bidirectional Encoder Representations from Transformers) model. Introduced by Facebook AI in 2019, RoBERTa is designed to enhance the pre-training of language representations by addressing some limitations and optimizing certain aspects of BERT. The key differences and improvements in RoBERTa include: Dynamic Masking:

Unlike BERT, which uses static masking during pre-training, RoBERTa employs dynamic masking. Larger Mini-batches and Learning Rates: RoBERTa uses larger mini-batches and learning rates during pre-training. No Sentence Order Prediction (SOP) Loss: BERT pre-training includes a task called Sentence Order Prediction, where the model learns to predict the order of two sentences. RoBERTa removes this task. More Training Data: RoBERTa is trained on a larger corpus compared to the original BERT model. Longer Training Duration: RoBERTa is trained for a more extended period. These modifications contribute to RoBERTa's robust performance across a variety of natural language processing tasks.

The pre-trained RoBERTa model we choose is downloaded from hugging face cardiffnlp/twitter-roberta-base-sentiment-latest[3]. This is a RoBERTa-base model trained on ~154M tweets from January 2018 to December 2022, and fine-tuned for sentiment analysis with the TweetEval benchmark. According to their paper, even for the original model, their performance on sentiment is better than base RoBERTa model.

We focus on the review and rating columns. We compared the number of positive rating and negative rating and found that ~78% of ratings are positive (20175) vs ~22 % ratings are negative (5646).



Train, valid and test data are randomly spit as 70%(18216), 20%(5113) and 10%(2582) of all the data. The distribution of positive and negative ratings is the same as the original data:

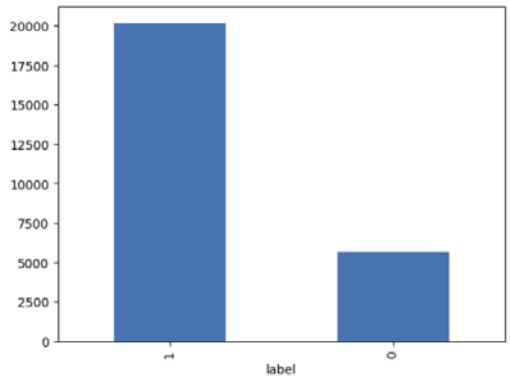
	Train	Valid	Test
Postive	14093	4015	2067
Negative	4033	1098	515

I used GPU to decrease the computational time significantly. Accuracy is selected as the metric to tuning the parameters. All the data are preprocessed to padding with max_length=128. The learning rate is chosen as 1e-5, weight_decay as 0.01. The model is trained for 10 epochs. The accuracy on test data set achieves 92%. The performance on positive and negative ratings in the test data are different.

label	**Precision**	**Recall**	**F1-score**

0	0.8	0.77	0.78
1	0.94	0.95	0.95

I further validate the mode on independent data set. After the same clean procedure and removing 1009 overlaps with original data set, we got 23609 reviews and corresponding ratings. I implemented the same classification on positive and negative ratings. The distribution of positive (18249) and negative (5360) ratings is similar to the original train data.



The performance on the independent dataset achieves 91% in accuracy. The performance on positive and negative ratings is also similar.

label	Precision	Recall	F1-score
1	0.84	0.74	0.78
0	0.93	0.96	0.94

Takeaways

Based on the above analysis and modeling, we can find that sentiment analysis can help the retailer to find out the key points to improve their service to promote customers experience. We also found that tf-idf encoding performs great in the analysis which expose much more informational features comparing to direct words count.

From the modeling section, we can found that with pre-trained RoBERTa model from tweet and fine-tuning with Walmart review data, we can have great performance in classifying customers reviews into positive and negative ratings. The accuracy achieved is around 91%. The model performs better on positive rating reviews which is as high as 95% in F1-score.

Future

Making use the customer review data to improve service is very meaningful and important. In the future, we can build a real-time application that will automatically update the review data and detect the key points to improve the customer service. We can also run the sentiment analysis with different

time range which will help us understanding customer's request changing and validating the outcome of improving customer service.

For model part, we find that there is still a big room to improve the performance on negative rating prediction. Although, we can use under sampling to have balanced positive and negative rating dataset, it is apparently not the real case. Since in the independent data set, we can find similar bias on positive ratings. Since most of the customers are rational, when they make a purchase, they probably have done some research and don't buy a product like flipping a coin. Therefore, the purchase itself is already a vote on positive rating. I need to find a method that can accommodate the biased input and achieve balanced well performance.

1. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805.
2. Liu Y, Ott M, Goyal N, Du J, Joshi M, et al. (2019) Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:190711692.
3. Loureiro D, Barbieri F, Neves L, Anke LE, Camacho-Collados J (2022) Timelms: Diachronic language models from twitter. arXiv preprint arXiv:220203829.