Text Mining of Customer Reviews on Masks (Group11 2021)

Cao Xiangyue (A0218943X) Huang Sixuan (A0049228B)
Xiong Zeyu (A0218858M) Xu Hao (A0218836W) Yu Zhe (A0218820J)

Abstract

Affected by the COVID-19 situation, the sales of masks on E-Commerce online shopping platforms grow fast. In this project, we propose a 3-stage methodology framework to classify the aspects of pain points behind the customers' negative reviews on masks. First, a method based on sentiment analysis is used to separate negative and positive reviews. Then, a similarity-based method helps to cluster the negative reviews related to different aspects, namely delivery, product and service. Finally, supervised machine learning models are developed to predict the aspect labels for the negative reviews. The proposed method is found to address the multi-label classification problem well, and can easily be extended to larger datasets, other products or other industries for drawing business insights from customer reviews.

1. Problem Statement

The E-Commerce market is expanding rapidly under the impact of COVID-19. According to the data from Statista, the number of users in Singapore is likely to grow from 3.1 million in 2020 to 4.1 million in 2025, and the revenue is expected to show an annual growth rate of 9.9%. Unlike traditional retailing services, there are limited accesses for E-Commerce platforms and sellers to approach every customer, which resulted in extreme difficulties for them to understand customers, attract potential sales, and reduce customer churn. Given the fact of great improvement in text mining techniques and computational powers, customer reviews become measurable and can play an essential role to allow sellers and platforms to identify problems and take actions accordingly.

1.1 Objective Scope

The objective of this study is to understand and classify the pain points behind customers' negative reviews, which would allow the E-commerce platform and the sellers to target specific problems and improve overall customer services.

To deepen our learning on the impact of text mining on customer reviews, our group decided to focus on mask customer reviews. Singapore government announced that it is mandatory to wear masks outside from April 14th, 2020. The sales of the mask on the E-Commerce platform grow rapidly since the mask becomes a necessity in daily life. The sudden sales increment also brings unprecedented customer reviews, which deserve a typical case study.

Besides, the analytical approach raised in this case study can be generalized to larger datasets and to other products, other E-commerce platforms, and even other industries, to improve customer services efficiently.

1.2 Overall Methodology

Three major analytic approaches have been raised to solve the problems in different stages throughout the case study. With scraped mask review data from Shopee, sentiment analysis, along with the rating information from customer reviews, is used in the first step to identify the negative reviews out of entire scraped dataset. Consequently, with specific pattern understanding from exploratory analysis, clustering analysis is conducted to automate the labelling and conclude the customer reviews from training dataset into relative aspects in the online shopping processes. Finally, classification analysis with various algorithms is applied for the prediction in the new incoming customer reviews. As a result, the entire methodology would enable an automate agile approach to identify analyze and classify every negative customer review into different aspects in the online shopping process, which will bring significant business impact to both E-Commerce platform and sellers.

2. Dataset Exploration

40,678 customer reviews from the top 18 mask sellers in Shopee are scraped. Even though the dataset is a bit small, the entire analytical methodology developed in this case will be applicable to any other larger datasets. Apart from the customer reviews, other information such as time, seller, customer, review rating, are collected for extended studies in this project.

Table 1 Variable Description

VARIABLE	DESCRIPTION
review_id	ID generated when scraping
time	Time of a customer left comment, in the format of "Year-Month-Date Hour-Minute"
seller_name	Name of the mask seller
preoduct_name	Name of selling masks from seller
review_content	Full content of a customer's review. Null if the customer does not leave any content.
customer_name	Some customers use full name while some customers' name has a format of "x*****x" if he/she leaves ananonymous review.
review_stars	Review rating from a customer in a range of 1 to 5, with 1 as the lowest rating.

2.1 Data Overview

All collected customer reviews are left from 2020 April to 2021 February. Across all 18 sellers' customer reviews, sellers with more amount of customer reviews do not necessarily receive more low-star complaints, which indicates that sellers' product and provided services do have certain differences in the entire shopping process, and customers' negative reviews can be an indicator to evaluate seller performances.

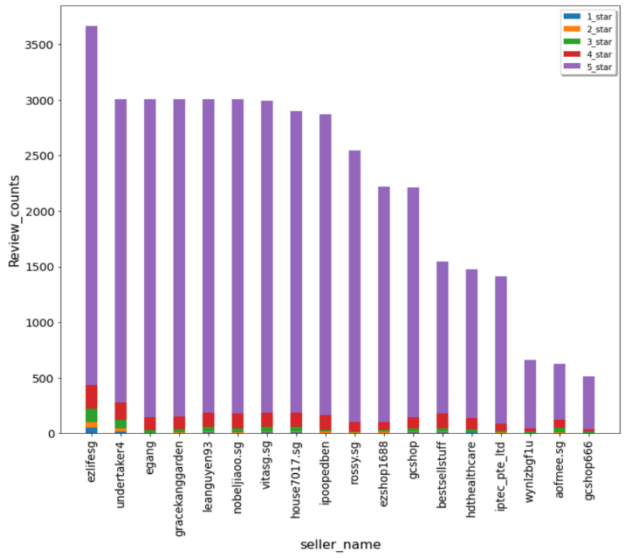

Figure 1. Number of Customer Reviews for Each Seller.

2.2 Negative Customer Reviews

There are around 90% of the review ratings are 5 stars, resulting in a limited space for collecting negative reviews. There is often a specific reason if a customer does not give full stars. Negative sentiments can be found easily in some high score (4 or 3 stars) reviews. Therefore, sentiment analysis is needed to identify the overall sentiment of a certain review.

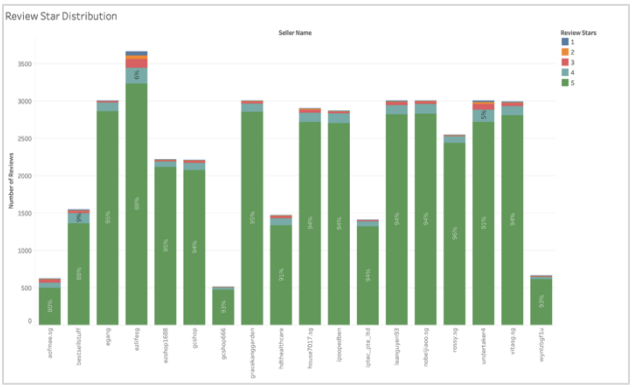

Figure 2. Review Star Distribution for Each Seller

2.3 Increment of Negative Reviews

As the selling amount improves, the percentages of low score 1 and 2 stars, are increasing which means customers are more sensitive to the online experience.

Therefore, there is a need to identify the aspects behind the negative reviews so that corresponding measures can be taken to overcome the problems timely.

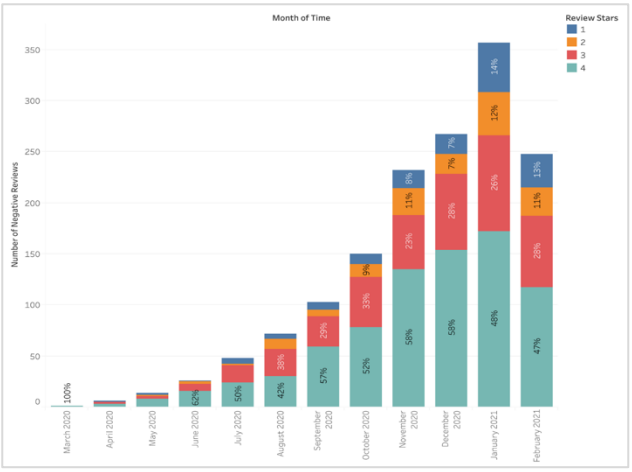

Figure 3. Number of Negative Reviews along Time

2.4 Reviews on Multiple Aspects

Negative reviews do not always complain about only one aspect. There are many reviews that indicate pain points from multiple aspects. For example:

'*Fast delivery.* ***Box dented. Quality is thin.*** *I tried to wear n the string broke.'*

'Ordered 2 bxs & received items in a plastic bag. ***One box is dented & opened*** *at the bottom when* ***the bag was unwrapped*** *. Bcos the masks r wrapped in a plastic bag,*

*the masks are still protected. Found the **mask not symmetrical & doesn't fit snugly at the chin** when worn. Had to tie a knot at one ear.'*

The first review illustrates the points from both delivery and quality, while the second review illustrates the pain points from delivery quality as well as services. Therefore, a multi-labeling approach is needed to identify all corresponding pain points behind every customer review to improve both sellers' and platforms' performances.

3. Methodology & Modeling Development

This section details the proposed 3-stage methodology which separates negative and positive customer reviews, and identifies the reviewing aspects of the negative reviews. A flowchart of the methodology main framework can be found in **Figure 4**.

The first stage involves separating negative and positive customer reviews using a sentiment-based method, as described in Section 3.1. At the second stage, a similarity-based method is used to automatically cluster the negative reviews into different aspects, as detailed in Section 3.2. Finally, Section 3.3 discusses the third stage, which develops supervised learning models to address the multi-label classification problem of identifying the pain points behind the negative reviews.

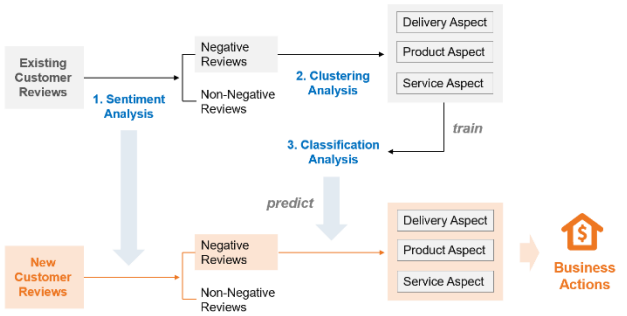

Figure 4. Flowchart of the proposed 3-stage method.

3.1 Sentiment Analysis

3.1.1 Problem Description

This section uses the web scraping raw data as input and aims to develop a method to automatically identify any negative sentiments contained in the customer reviews. With the proposed method, the work of manual intervention can be reduced on separating negative and positive reviews.

There are mainly various challenges for this task. Firstly, we need to recognize any negative sentiment that exists in high score reviews whose overall sentiment might be actual positive. Besides, in the raw data, there is only a rating star for a whole customer review, while no sentiment label is available for either the whole sentence or any of the sub-sentences. Therefore, we will need to properly determine the target labels for developing a classification model to separate the negative and positive sentiments.

3.1.2 Data Processing

Before analysis, text cleaning and pre-processing are carried out, mainly including the following aspects.

- Removal of null data and white spaces.
- Emojis are converted to strings as they may convey certain sentiments. For example, 😊 is converted to "smiling face".
- Word lemmatization is done to reduce the different forms of a word to one single form.
- Sentence tokenization is performed to split each sentence into sub-sentences. Sentences are split by punctuations and words such as 'however', 'but' so that different sentiments can be captured by different sub-sentences.

3.1.3 Analysis Methodology

To identify any negative sentiments and separate the customer reviews into negative and positive groups, it is found that pure sentiment analysis on the whole sentences does not give satisfactory results. Therefore, a method based on the combination of sentiment analysis and supervised classification model is proposed in this section, with details described below.

1. Training Data Selection. Through observation, it is noticed that the 1-3 star reviews, with a total number of 760, almost all contain negative sentiments. Therefore, they are all included in the training dataset to be used as samples with the 'negative' label. Moreover, as the 5-star rating reviews rarely contain negative sentiment, 800 of them are randomly sampled and included in the training dataset as 'positive' samples. Thus, a balanced training dataset is obtained. Since the 4-star reviews are highly mixed of positive and negative comments, such as "fast delivery, box dented", they are not included in the training dataset.

2. Sentiment Analysis. To 'clean' the sentiment in each customer review in the training dataset, sentiment analysis is carried out for each sub-sentence using sentence tokenization and SentimentIntensityAnalyer from the nltk package. For the 1-3 star reviews with the 'negative' label, any sub-sentences with positive sentiment scores are removed. On the other hand, for the 5-star reviews with the 'positive' label, any sub-sentences with negative sentiment scores are excluded. After this, we obtain a training dataset with each sample only contains clean 'negative' or 'positive' sentiment, and we have transformed

the rating stars to sentiment target labels of 'negative' and 'positive'.

3. Model Development. Due to their high performances in most text classification tasks and high efficiency, CountVectorizer and Multinominal Naïve Bayes models are used for the sentiment classification task. 5-fold cross-validation and heuristic Bayes Optimization are used to determine the optimized hyperparameters (min_df=2 and alpha= 0.70344)

3.1.4 RESULTS DISCUSSION

The classification model obtained in **Section 3.1.3** is used to predict the sentiment of all 1-4 star customer reviews in the original dataset. 5-star reviews are assumed to be all positive, so they are excluded in the following stages of this study. The sentences predicted as 'positive' are filtered out and only sentences predicted as 'negative' are collected for determining the reviewing aspects in the next stage. Before that, manual confirmation and adjustments on the predicted labels are done for those customer reviews with 'blurred' prediction probability.

The prediction results of the proposed method (customized sentiment classification model) are evaluated in **Table 2**, by comparing to the true sentiment labels obtained after manual adjustments. Meanwhile, the prediction results based on solely sentiment analysis of the whole customer reviews are also shown in **Table 2** for comparison. A general conclusion is that the performance of the customized sentiment classification model improves significantly compared to the built-in classifier in the *ntlk* package, enabling less manual work on separating negative and positive reviews. Meanwhile, high accuracy is achieved with a small amount of training data, meaning that this method can also be efficient for large datasets.

Table 2. Prediction Results of Sentiment Analysis

	Class	Precision	Recall	F1-score
Nltk Sentiment package prediction	negative	0.97	0.50	0.66
	positive	0.41	0.96	0.58
	Accuracy			0.62
	Macro avg	0.69	0.73	0.62
	Weighted avg	0.90	0.89	0.89
Model prediction	negative	0.96	0.89	0.93
	positive	0.68	0.87	0.76
	Accuracy			0.89
	Macro avg	0.82	0.88	0.85
	Weighted avg	0.83	0.62	0.64

Finally, after the separation of the negative and positive customer reviews, it is easier to identify the related aspects behind the negative reviews. As can be seen in **Figure 5**, a word cloud based on the negative customer reviews shows that there are mainly 3 aspects in concern:

1. Delivery, involving slow delivery, wrong delivery and dented packaging box, etc.
2. Product, mainly involving the quality problems of the masks, such as smelly masks, thin masks and loose strings, etc.
3. Service. This includes no response or no refund from the seller, wrong item sent by the seller, etc.

Analysis at later stages will be based on these three aspects found in this section.

Figure 5. Word Cloud of All Negative Reviews Obtained after Sentiment Analysis Stage.

3.2 Unsupervised Machine Learning

3.2.1 PROBLEM DESCRIPTION

With the negative customer review dataset obtained from the analysis in **Section 3.1**, this section aims to automatically cluster these negative reviews to their corresponding clusters, including delivery, product and service. It is hoped that the outcome at this stage could reduce the manual work of aspect label assignments.

The conventional KMeans clustering method tends to generate similar amounts of samples in each cluster. However, an overview of the data in this project indicates that the numbers of negative reviews in the three aspects are quite different. On the other hand, text clustering is a challenging task because it is difficult for the algorithms to capture the word context and the word sequences. There can also be mis-spelling or abbreviation which make the text unrecognizable. Therefore, to improve the outcome of text clustering in this project, we try to propose an innovative similarity-based clustering method, and will compare its performance to the KMeans method.

3.2.2 ANALYSIS METHODOLOGY

The proposed similarity-based clustering method mainly consists of three steps as detailed below.

1. Vectorization. First, the customer reviews obtained from the last stage, which include negative sentiment to different extents, are vectorized. Different vectorization approaches such as TF-IDF, CountVectorizer, Word2Vec or

BERT can be used. In this project, the start-of-the-art BERT model is used as it gives the best results. To achieve better clustering results, we should not use the whole sentence for vectorization. Using the sentiment analysis method in **Section 3.1**, any sub-sentences containing positive sentiment are removed before vectorization of the customer reviews. The output vector is obtained by taking an average of the vectors extracted from the last layer of the BertModel.

4. Initialization of Cluster Centroids. The cluster number of k=3 (delivery, product and service) is predetermined based on the analysis results in **Section 4.1** and confirmed by the clustering performances. To initialize the cluster centroids, aspect-based keywords for each cluster as shown in **Table 3** are used. For each negative customer review, if it consists of any of the aspect-based keywords, it is included in the initial 'core data' for the corresponding cluster. Afterwards, the centroid of each cluster is determined by taking an average of the BERT vectors of the 'core data'.

2. Similarity-based Clustering. Various similarity metrics including Euclidean Distance, Cosine Similarity or Pearson Similarity can be considered. In this study, the cosine similarity of each negative review to each of the cluster centroids is computed. After that, each review is assigned to the cluster that is closest to it.

Two cases with different aspect-based keyword lists are studied for the proposed method here. For the Base Case, only the most direct aspect-related word is included in the keyword list. This represents the simplest scenario and it helps to evaluate the feasibility of extending this method to large-scale datasets. In the second case, namely the Improved Case, we try to add more keywords to each list, which undermines the commuting efficiency but improves the clustering performance.

Table 3. Keyword lists for the Base Case and the Improved Case. Base Case only considers the most aspect-related word, while more keywords are added to the Improved Case to improve clustering results.

Case	Keyword List
Base case	Delivery = ['delivery'] Product = ['quality'] Service = ['service']
Improved case	delivery = ['box', 'date', 'days', 'delivery', 'dented', 'received', 'plastic', 'sealed', 'time', 'week'] product = ['2nd', 'big', 'black', 'different', 'easily', 'loose', 'material', 'nose', 'product', 'quality', 'size', 'small', 'smell', 'soft', 'surgical', 'thick', 'thin', 'tight'] service = ['service']

3.2.3 Results Discussions

It is noted that a minority of the negative customer reviews could involve more than one aspect, but the clustering method only assigns one aspect label to each of the reviews. After the clustering analysis, manual efforts are taken to adjust those 'blurred' customer reviews that have close similarities to different cluster centroids, to check their labels and assign multiple labels if needed.

The result evaluations in this section are based on the clustered labels and the true labels. Results of the KMeans case is also included for comparison. For the KMeans case, the sentence vectors obtained in Step 1 of Section 4.2.2 are used for clustering with cluster number set as k=3.

Table 4. Results Comparison among Different Methods.

Aspect	Score	Kmeans	Base Case	Improved Case
Delivery	Macro f1	0.53	0.54	0.64
	Accuracy	0.64	0.58	0.73
Product	Macro f1	0.48	0.65	0.69
	Accuracy	0.49	0.65	0.72
Service	Macro f1	0.55	0.64	0.65
	Accuracy	0.67	0.81	0.79

From **Table 4**, it can be seen that the proposed clustering method outperforms the Kmeans method. One reason is that KMeans tends to have a similar amount of data in each cluster, but the numbers of negative reviews involving various aspects are different in this case (see **Figure 5**). Another possible reason is that the features in the sentence vectors are not distinct enough to define the cluster boundaries, and this explains why using a keyword list helps to improve the results.

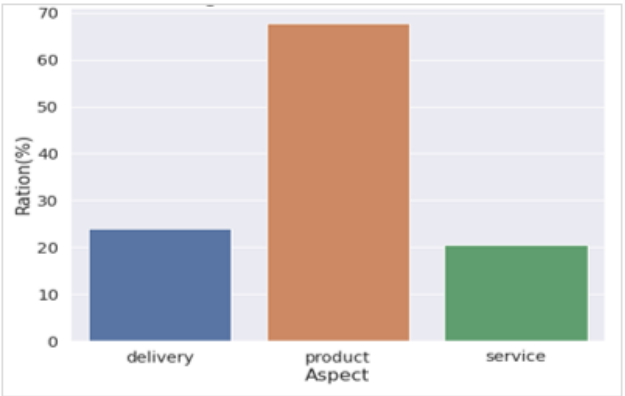

Figure 5. Distribution of negative customer reviews replated to the three different aspects of delivery, product and service.

Another important observation from the results is that, even with only a single word in the keyword list, the Base Case gets close scores to the Improved Case. This means

that a simple initial setting of the method can achieve results with high enough accuracy, showing the potential of this method to be also used for larger datasets.

The Improved Case has the best performance among the three cases. It shows that by adding more appropriate words to the keyword lists, the algorithm can better identify the cluster centroid, thus improves the clustering outcome. Therefore, one possible way to improve the current method is keeping updating the keyword lists for each aspect, so that higher quality data can be included in the initial 'core data' for each cluster. Besides, it is possible to iteratively update the cluster centroids as data are added to each cluster. However, trade-offs between computing efficiency and computing accuracy would need to be made if the measures mentioned above were to be taken.

3.2.4 Results Visualization

The effects of KMeans clustering and Improve Case are shown in **Figure 6**, where the distances of each customer review to the three cluster centroids are mapped into a two-dimensional space. The True Label Case (manual adjustment of 'blurred' labels after clustering) is also shown for comparison. But it should be noted that for better visualization effect, the multi-label data are excluded in the plot for the True Label Case. **Figure 6** again confirms that the Improved Case obviously outperforms the KMeans case.

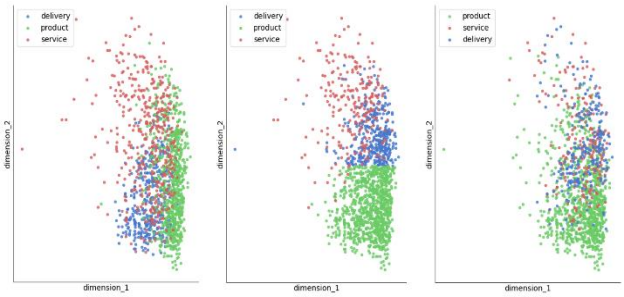

Figure 6. Clustering visualization of Kmeans (left), Proposed Improved Case (middle) and True Labels (right).

With the outcomes of the aspect-based labels for each negative customer review at this stage, we are able to clearly understand the pain points behind each aspect, as shown in the word clouds in **Figure 7**.

Figure 7. Word clouds showing paint points behind the negative customer reviews for the aspects of delivery, product and service.

3.3 Supervised Machine Learning

3.3.1 Problem Description

In this section, we leverage the sentence vectors extracted from BERT and various classic machine learning and deep learning models to identify the aspects of the complaints in negative customer reviews and classify them into the corresponding categories. We intend to judge whether a negative review is related to each of the three aspects, so the multi-label task can then be divided into three binary classification problems.

Afterwards, the prediction performances of the classifiers would be evaluated comprehensively based on AUC scores and f1-macro scores of the three aspects. It is expected that the best model can automatically predict the related aspects of the future incoming comments accurately.

3.3.2 Methodology

The main workflow is shown in **Figure 8**. The first step is to input review text data into a pre-trained BERT encoder considering that BERT is excel at capturing sentence patterns and semantic meanings of text data and acquire a matrix of word vectors for each comment. In the second step, these vectors serve as features and are input into the models to accomplish the three binary classification tasks with respect to delivery, product, and service. For this study, we choose Naïve Bayes and K-Nearest Neighbors as benchmark models, and then employ neural network models including CNN, RNN, Bidirectional LSTM, and BERT classifier to compare their performances on the test set.

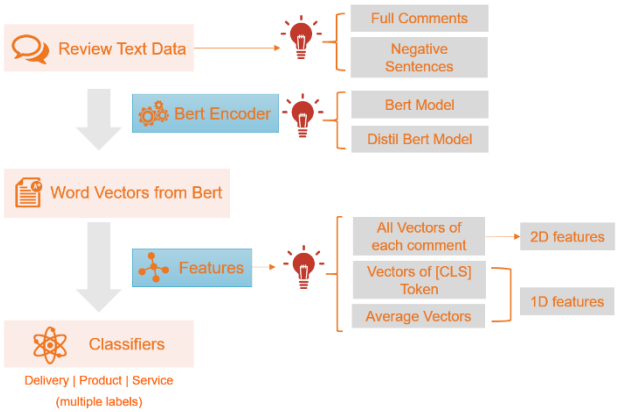

Figure 8. Framework of Supervised Learning

The steps of the task can be further divided into different cases according to the options of (1) input text data, (2) encoder, and (3) the form of features.

1. Firstly, when putting text data into BERT encoder, we can choose to input the full original comments, or delete positive sentences and only input negative ones to remove noises. We suppose that positive sentences containing words related to delivery, product, and service could easily confuse the model and could pose

an extra burden to require the model to distinguish between praises and complaints.

2. As for the encoder, besides the original BERT layer, there is also another variation called DistilBERT, which is a lighter and faster version of BERT that roughly retains the performance. We will try both encoders to see which would give a better result.
3. Furthermore, in terms of the form of features, we can train the models with a two-dimensional matrix consisting of all the word vectors of each comment obtained from BERT. On the other hand, we can choose to only input the vector of the first token [CLS]. According to the original BERT paper, the final hidden state corresponding to the [CLS] token can be considered as an aggregate sequence representation for a classification task. Alternatively, the average vector over all the vectors is also applicable. In this way, the feature value would be a one-dimensional sequence.

Naive Bayes and KNN are used as benchmark models to simply examine how well BERT vectors can capture the patterns in sentences and contribute to the classification in this study. Afterwards, we implement Bidirectional LSTM as a baseline model over all the options to select the case with the best performance and apply the selected case to other neural networks for model evaluation and selection.

3.3.3 Model Evaluation

Table 5 displays the prediction results on test data of Bidirectional LSTM in 6 cases respectively. Model performances are evaluated based on AUC and f1-macro score. Cases with BERT encoder are not shown here because Distil BERT encoder outperforms in all the cases. The outcomes from Bidirectional LSTM demonstrate that the cases with average vectors of or all vectors of negative sentences as input and Distil BERT as encoder layer will bring the best performance. So, these two cases are applied to CNN, RNN, as well as BERT classifier, to further evaluate which model with which case would be the best. In terms of three labels, prediction performance on delivery and quality are typically better than that of service, probably resulting from the imbalanced dataset, where the comments pointing to the former two problems comprise a larger proportion.

From **Table 6** of the final model selection results, we can see that the NB and KNN can already achieve a quite good outcome. Lastly, fine-tuned BERT classifier trained on negative sentences outperforms the others. It is probably because the Multi-head Attention mechanism of the BERT layer in the encoder can better capture semantic meanings and similarity of sentences within the same topic. Another reason could be that, in this case, the model is trained directly on customers' comments so the weights in BERT layers are fine-tuned with new text data in our dataset based on the initial pre-trained model.

Table 5. Prediction Results of Bidirectional LSTM over 6 Cases.

Options	Cases					
	1	2	3	4	5	6
Feature	[CLS]	AVG	2D	[CLS]	AVG	2D
Sentence	Full	Full	Full	Nega-tive	Nega-tive	Nega-tive
Encoder	Distil Bert	Distil Bert	Distil Bert	Distil Bert	Distil Bert	Distil Bert
F1-macro	0.791	0.805	0.813	0.804	0.838	0.846
(del.	0.823	0.851	0.834	0.830	0.872	0.836
pro.	0.779	0.784	0.747	0.723	0.761	0.749
ser.)						
AUC	0.899	0.908	0.915	0.896	0.932	0.931
(del.	0.909	0.925	0.924	0.916	0.929	0.923
pro.	0.846	0.877	0.867	0.846	0.844	0.831
ser.)						

Table 6. Prediction Results for Model Selection.

Options	NB	KNN	LSTM	CNN	RNN	BERT
Feature	AVG	AVG	AVG	2D	2D	/
Sentence	Negative					
Encoder	Distil Bert	Distil Bert	Distil Bert	Distil Bert	Distil Bert	Fine-tuned Bert
F1	0.735	0.731	0.838	0.873	0.802	0.873
(del.	0.781	0.808	0.872	0.877	0.854	0.886
pro.	0.682	0.692	0.761	0.773	0.741	0.803
ser.)						
AUC	0.843	0.866	0.932	0.933	0.897	0.943
(del.	0.878	0.911	0.929	0.933	0.911	0.941
pro.	0.797	0.832	0.844	0.869	0.816	0.850
ser.)						

Overall, the classification of customers' comments is challenging. Many instances are difficult to be classified even manually because of the complex context, ambiguous attitude, and fuzzy aspects in the real-world reviews. Three comments are taken as examples to illustrate these three difficulties respectively.

1. Complex context:

   'Dont understand the label in pic 1. 19083 is a surgical mask standard, yet implement is GB/T32610 which is a nonmedical mask? Mask look genuine, but the standard is confusing.'

   The buyer was doubting information authenticity of the mask product and mentioned details about the code

label indicating the type of mask in the comment. However, it is quite hard to classify it into 'product' class since this information is rare in the corpus. Besides, there are typically numerous different contexts with highly specific information in the dataset.

2. Ambiguous attitude:

   'Seller attached a notice saying they are sorry because the stocks not available due to delay. And also a small gift. Anyway, the masks I received is ok too. But I realized later that it is Children Mask that I received.'

   Although thr quality of the mask was acceptable, and the overall attitude was neutral, actually the product type that the buyer received was even wrong, which was a big mistake. Nonetheless, the buyer did not show much sentiment, so it is difficult to classify it as positive or negative.

3. Fuzzy aspects

   'Bought 4 boxes. Material is fine, value for money. But buyer need to double check the packaging if they are packed and sealed properly before storage. Out of 3 boxes, only 1 is properly sealed and the box is flattened.'

   This comment complained that the product was not properly sealed, and the package was damaged. But the buyer did not say explicitly whether the seller's packaging service or delivery should be blamed.

In summary, the current models have achieved acceptable performance despite of tolerable errors. Misclassification is understandable in some cases and there is room for further improvements. Possible improvements include: (1) training the classifier with a larger corpus to identify more detailed aspects of complaints; (2) adding manual features to help the model identify context-specific representations.

4 Extended Studies

As discussed before, the key motivation to analyze the reviews of the mask is that it has become a necessity given the covid pandemic. Therefore, it could be interesting to explore the relationship between covid relevant data and customers' sentiments. In addition, given the analysis from previous sections, it could be insightful to delve deeper into customers' sentiments under the three aspects of the negative reviews, and see how they change over time.

4.1 COVID Cases and Sentiment Score

In order to reflect the overall sentiment of the customers, the average sentiment score by month is calculated from all reviews (1-5 stars). Based on **Figure 8**, there exists a decreasing trend in customers' sentiment over time in general. One thing to note is that there is a lag between a customer's purchasing time and reviewing time. To explore the relationship between sentiment score and covid severeness, the number of covid confirmed cases by month is also plotted. Although covid cases start to fall since May 2020, the average sentiment score continues to drop. One possible reason can be that the mandatory mask-wearing policy has not been relaxed, so customers are still demanding.

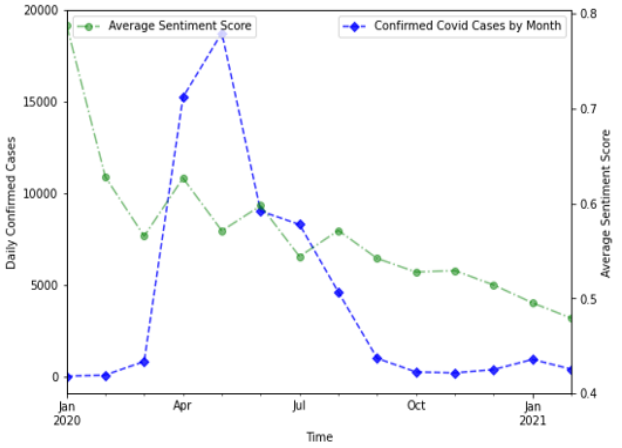

Figure 8. Covid Cases & Avg. Sentiment Score by Month

4.2 Aspect-based Sentiment Score

According to the outcome of previous analyses, aspect labels are available for all negative reviews. Analyzing the change of average sentiment score with respect to each aspect over time could also produce meaningful insights. Referring to Fig xx, sentiment scores are very low, mostly negative, among all aspects during the peak time of covid cases (From April 2020 to July 2020). Since Aug 2020, as the covid severeness has been mitigated and the number of confirmed cases continues to fall, the scores for all three aspects gradually improve.

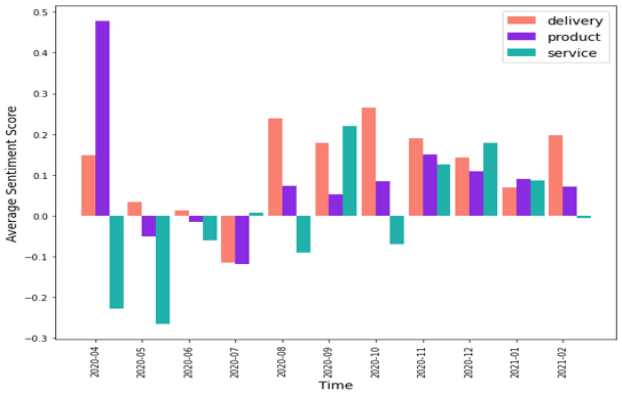

Figure 9. Aspect-based Avg. Sentiment Score by Month

To further understand how the changes of sentiment score within each aspect may be affected by the severeness of covid, the number of covid cases by month is plotted within each aspect.

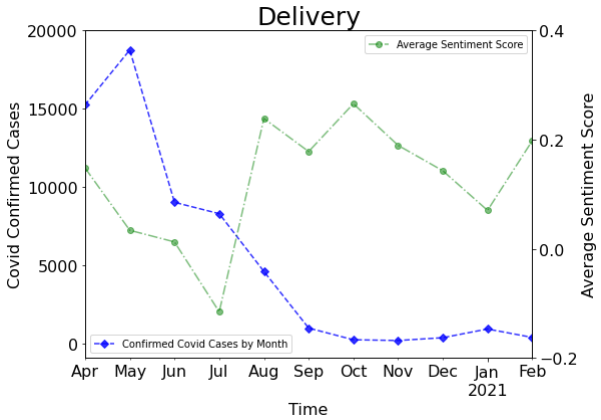

Figure 10. Covid Cases & Delivery Avg. Sentiment Score

For the negative reviews of the delivery aspect**,** initially, the sentiment score drops as the number of covid cases remains high. One possible reason is that when the government enforced mask-wearing, customers hoped to receive the masks fast, in other words, they are more demanding on the delivery perspective. The average score gradually improves as covid cases fall.

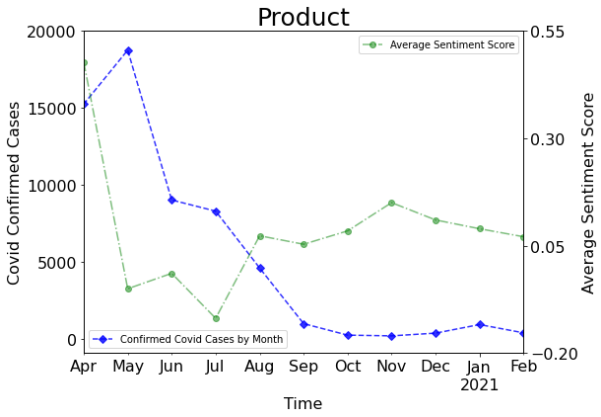

Figure 11. Covid Cases & Product Avg. Sentiment Score

For the product aspect, a huge drop when the covid cases reached the peak is observed. Another interesting finding is that the average score is still relatively low, fluctuating around 0.05, even when the number of cases decreased, which may indicate that customers remain rigorous on the product aspect (eg. Quality of the mask).

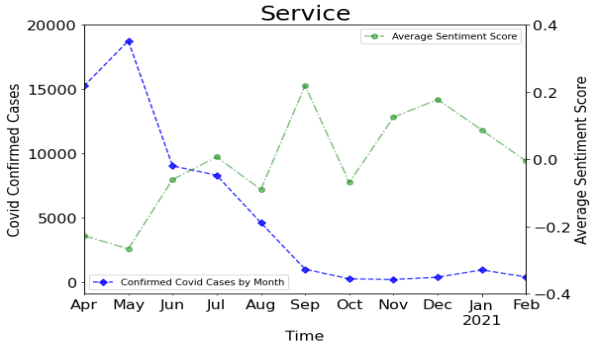

Figure 12. Covid Cases & Service Avg. Sentiment Score

Finally, for the service aspect, there was an initial decrease in the average score as the pandemic became more severe. However, it is observed that the average sentiment score shows an increasing trend along time in general. A possible explanation is that the platform and the sellers improved their services according to customers' negative reviews.

The last analysis to be carried out in this section is to examine how the frequent words in negative reviews change with time. As a result, the word cloud of negative reviews by month is generated. Due to the limitation of space, only three representative word clouds are demonstrated here.

Figure 13. Word Cloud of Negative Reviews by Month

In the early stage of the pandemic, which is from April 2020 to August 2020, the word clouds are quite similar, largely dominated by words like 'delivery', 'packing', and 'box', implying that the customers focused more on the delivery aspect back then. Since September 2020, the keyword 'quality' gradually came into play, which is quite reasonable. As customers have already experienced mask-wearing for a few months, they placed more emphasis on the product aspect, valuing more on issues like whether the mask has high quality and whether the mask is comfortable, etc. The last word cloud is from Feb 2021, showing that now the customers focus largely on both the delivery and the product aspect, which is consistent with the previous discussions.

5 Conclusion

Customer reviews can play an essential role to allow sellers and E-Commerce platforms to identify problems and take actions accordingly. With the observation that masks have become daily necessities under the COVID-19 situation, we collect customer reviews of mask selling from an E-Commerce platform and carry out analysis with the aim of classifying the paint points behind the negative comments. A 3-stage workflow (**Figure 4**) is proposed to address the multi-label classification problem, so that the customers' negative sentiments can be recognized and be related to different aspects of delivery, product or service.

At the first stage, as detailed in **Section 3.1**, a sentiment-based classification method is used to filter out the customer reviews containing any negative sentiment. It is found that the proposed method outperforms the method with pure sentiment analysis, enabling less manual work to be carried out on separating negative and positive reviews. Furthermore, the proposed method achieves relatively high

accuracy with a small amount of training data, which means it can also be efficient for use in large datasets.

The second stage, as described in **Section 3.2**, proposes a clustering method based on similarity calculation to automatically cluster the negative reviews into different aspects. The proposed method outperforms KMeans, which has difficulty at handling the imbalanced dataset. By using the aspect-based keyword list, the proposed method gets a set of high-quality initial 'core data' to determine the cluster centroids. It can achieve reasonably good results with only a single word as a start and can be improved by including more appropriate words in the keyword list.

The final stage discussed in **Section 3.3** addresses a multi-label classification problem. Supervised machine learning models are developed for the prediction of aspect labels for future customer reviews. A fine-tuned BERT Classifier using negative sub-sentences of the customer reviews shows the most outstanding performance. This is because the multi-head attention mechanism of the BERT layer can better capture semantic meanings and similarity, and the model weights in BERT layers are fine-tuned with new data based on the pre-trained model.

The negative customer reviews with aspect-based labels obtained from the analysis can be used to draw various business insights. Moreover, they can be used to compare the severeness of different aspects along time, so that decisions can be made on the priority of problem-solving.

There are a few potential improvements in this project. Because negative reviews are the minority among customer reviews, the data scale in this study is relatively small. Therefore, reviews from multiple platforms or different types of products can be included to obtain a larger dataset. In terms of methodology, the unsupervised machine learning can be improved by iteratively updating each cluster's centroid as data are added in; for the supervised ML, manual features can be added to help identify context-specific representations. Lastly, for models, more comprehensive parameter optimization can be applied, and more state-of-the-art models can be included to improve the prediction performances.

In conclusion, it is believed that the developed methodology framework of this project is applicable to larger datasets with trade-off considerations on computation efficiency and accuracy. Furthermore, the analytic approach can also be widely applied to other products or other industries to improve customer services efficiently.

Source Codes

https://github.com/justdropby/BT5153_Group11

References

eCommerce – Singapore | Statista Market Forecast. Retrieved from https://www.statista.com/outlook/243/124/ecommerce/singapore

The Map of E-commerce in Singapore – iPrice Singapore. Retrieved from https://iprice.sg/insights/mapofecommerce/

Joachims, T. Text categorization with support vector machines: Learning with many relevant features. Paper presented at the, 1398 137-142. Doi 1998:10.1007/s13928716

Gao, Fei, and Xuanming Su. "Manufacturing & Service Operations Management." Manufacturing & Service Operations Management 19, no. 1 2017: 84-98.

Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. the Journal of machine Learning research, 2003, 3: 993-1022.

Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv preprint arXiv:1706.03762, 2017.

Choi Y, Lee Y, Cho J, et al. Towards an appropriate query, key, and value computation for knowledge tracing[C]// Proceedings of the Seventh ACM Conference on Learning@ Scale. 2020: 341-344.

Charles9n. (2019, August 11). Charles9n/bert-sklearn. Retrieved April 25, 2021, from https://github.com/charles9n/bert-sklearn