Text Mining of Customer Reviews on Masks

BT 5153 Group Project, Group 11, Apr 2021

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

Agenda

01 **Introduction**
Object | Scope | Methodology

02 **Data Description**
Web Scraping | EDA

03 **Sentiment Analysis**
Separate Positive / Negative Reviews

04 **Unsupervised ML**
Aspect-based Review Clustering

05 **Supervised ML**
Prediction of Reviewing Aspects

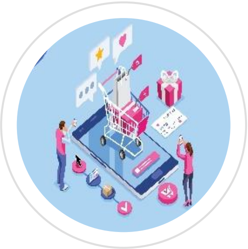
06 **Extended Studies**
Insights behind Customer Reviews

07 **Summary**
Highlights | Possible Improvements

01

Introduction

Motivation



Why E-Commerce

E-commerce market is expanding rapidly.

"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 (CAGR 2021-2025) of 9.9%."

Why Reviews

Customer reviews play an essential role.

E-commerce companies, and the sellers within the platforms, strive to attract new customers and reduce customer churn.

Customer reviews can help the companies to identify existing problems and take actions accordingly.

Why Mask

Online mask selling grows fast under COVID19

On April 14th, 2020, Singapore government announced that it is mandatory to wear masks when leaving the house.

The sales of mask on online platforms grows rapidly, as mask becomes a necessity in daily life. Its customer reviews deserve great attention.

Project Objective and Scope

Objective

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.

Significance

The analytic approach can be generalized to larger datasets and to other products, other E-commerce platforms, and even other industries, to improve customer services efficiently.

01

Sentiment Analysis

To separate positive and negative customer reviews automatically.

02

Unsupervised ML

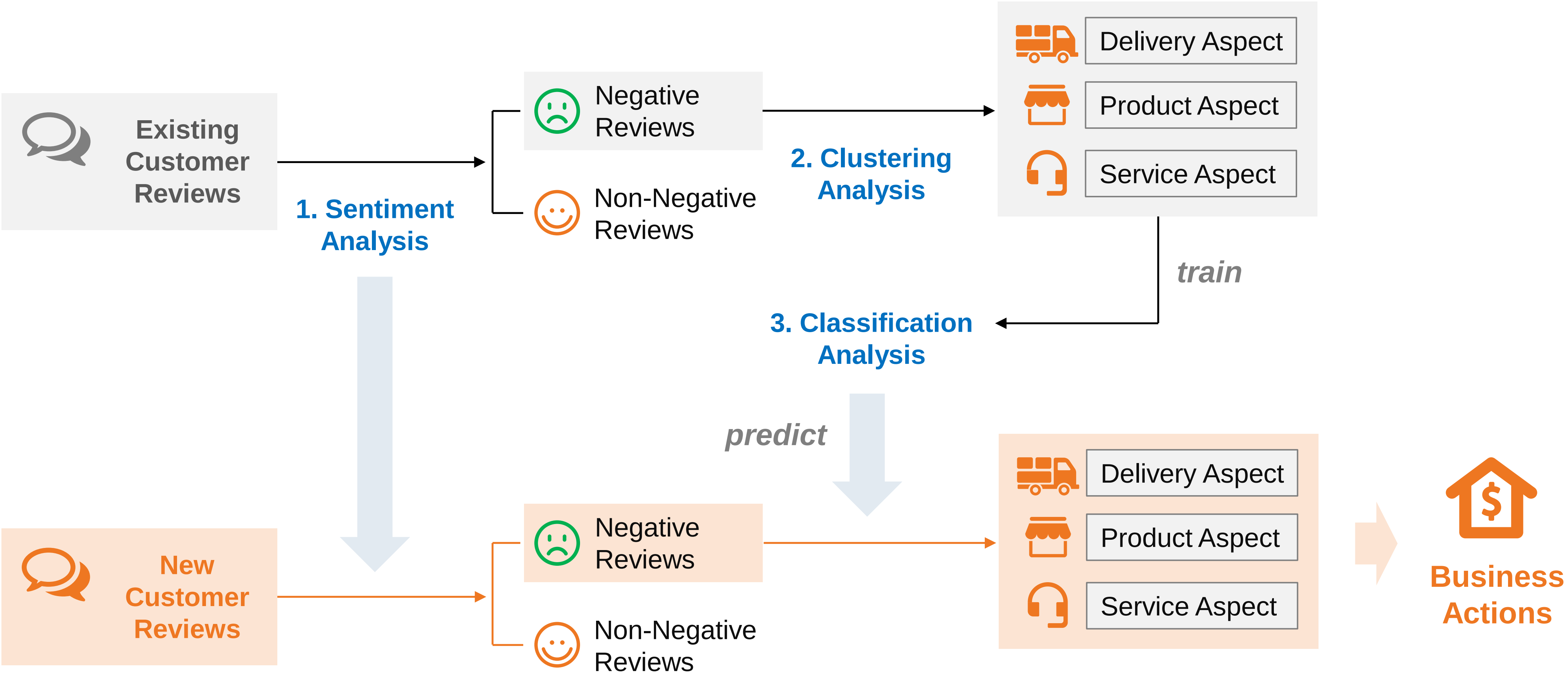
To cluster different negative reviews with respect to their reviewing aspects.

03

Supervised ML

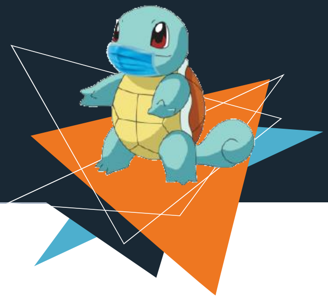
To classify future negative customer reviews into corresponding aspects.

Methodology Flowchart



02 Data Description

Data Collection



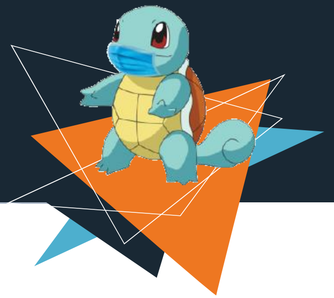
Web Scraping Data

- The online customer reviews are drawn from Shopee’s web pages.
- Top **18** mask sellers with **40,678** customer reviews of 1-5 stars.
- A relatively small dataset. But to keep in mind that the methodology developed in this study can be generalized to larger datasets.

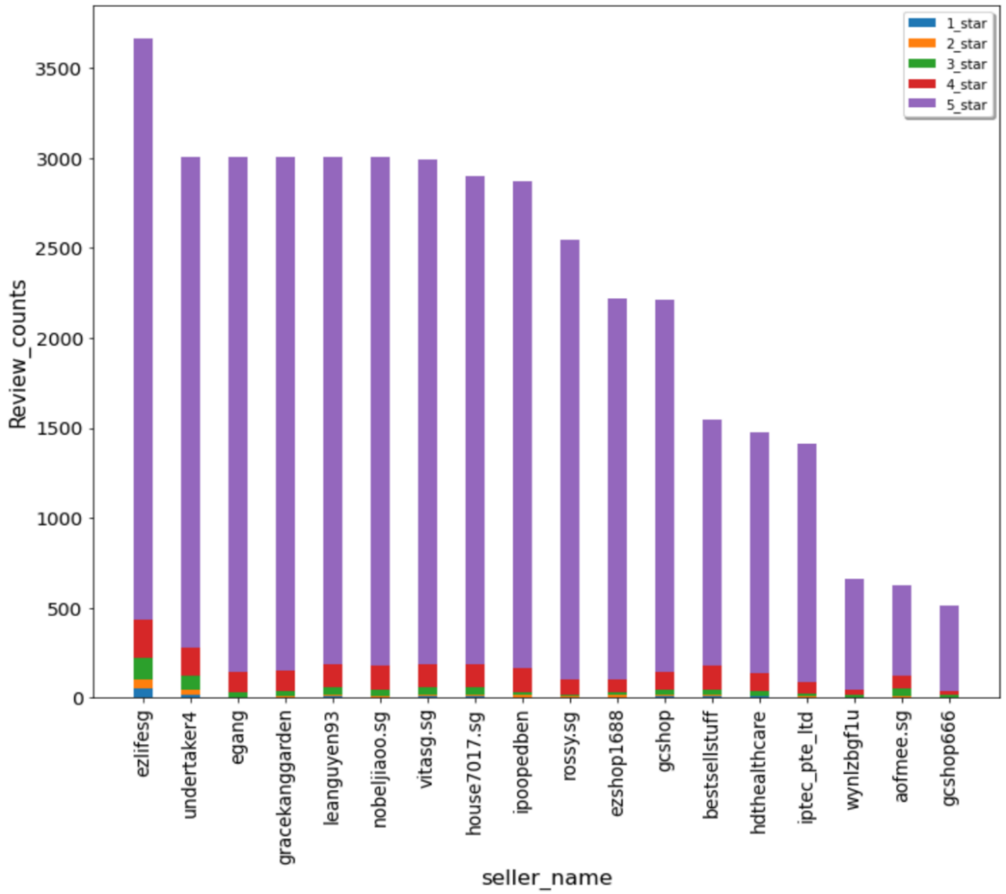
Examples of Web Scaping Data (Selected Columns)

review_id	time	seller_name	product_name	product_variation	review_content	customer_name_full	customer_name_anonymous	review_stars
2	2021-02-17 00:28:00	ezlifesg	Local Ready Stock 3 PLY Disposable Face Masks ...	Adult Black	Ordered on 6th, rec'd parcel on 10 Feb at 11.3...	NaN	w*****4	3
12	2021-02-08 21:23:00	ezlifesg	Local Ready Stock 3 PLY Disposable Face Masks ...	Adult Black	Fast delivery. Box dented. Quality is thin. I ...	NaN	n*****4	3
13	2021-02-12 11:39:00	ezlifesg	Local Ready Stock 3 PLY Disposable Face Masks ...	Adult Black	Second purchase the mask.. so disappointed not...	NaN	s*****l	2
14	2021-02-16 16:38:00	ezlifesg	Local Ready Stock 3 PLY Disposable Face Masks ...	Adult White	Received within a few days. However it is not ...	NaN	h*****i	3

EDA: Data Overview

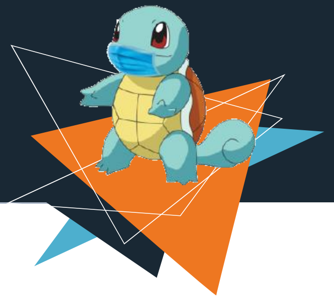


Sellers with more reviews don't necessarily receive more complaints, showing that the **customers' negative reviews can be an indicator of selling performances.**



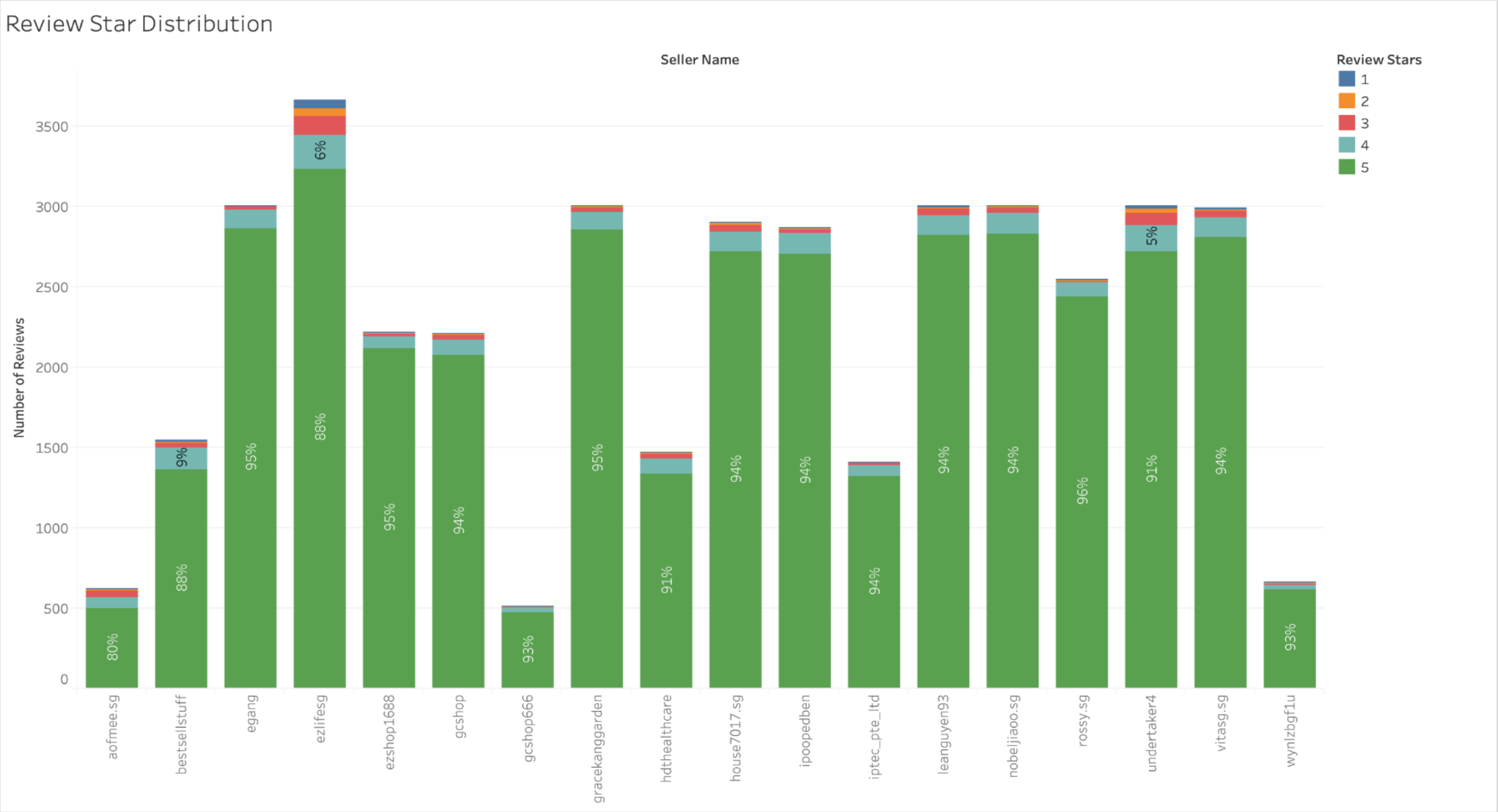
seller_name	Time Range	Number of Reviews					
		1 star	2 stars	3 stars	4 stars	5 stars	Total
aofmee.sg	2020-11~2021-02	6	6	39	75	497	623
bestsellstuff	2020-07~2021-02	9	8	26	139	1363	1545
egang	2020-04~2021-02	2	2	25	113	2864	3006
ezlifesg	2020-08~2021-02	56	49	115	212	3234	3666
ezshop1688	2020-07~2021-02	5	10	18	73	2116	2222
gcshop	2020-07~2021-02	9	7	27	99	2074	2216
gcshop666	2020-11~2021-02	2	2	10	23	475	512
gracekanggarden	2020-09~2021-02	4	7	30	108	2857	3006
hdthealthcare	2020-07~2021-02	10	3	28	94	1339	1474
house7017.sg	2020-12~2021-02	7	9	42	126	2718	2902
ipoopedben	2020-10~2021-02	4	10	21	129	2706	2870
iptec_pte_ltd	2020-09~2021-02	2	6	14	67	1325	1414
leanguyen93	2020-04~2021-02	11	8	43	123	2821	3006
nobeljiaoo.sg	2020-09~2021-02	6	4	35	135	2826	3006
rossy.sg	2020-01~2021-02	4	4	10	88	2441	2547
undertaker4	2020-06~2021-02	21	24	80	158	2723	3006
vitasg.sg	2020-04~2021-02	11	10	41	123	2809	2994
wynlzbgf1u	2020-11~2021-02	4	2	11	28	618	663
	total counts	**173**	**171**	**615**	**1913**	**37806**	**40678**

EDA: Imbalanced Data



There is a need to identify negative points even if the overall sentiment of a review is positive.

- About 90% of the ratings levels are 5 stars, leaving a very small room for collecting negative reviews.
- However, there is often a specific reason if a customer does not give full stars. Negative sentiments can be found in high score (4 or 3 stars) reviews.

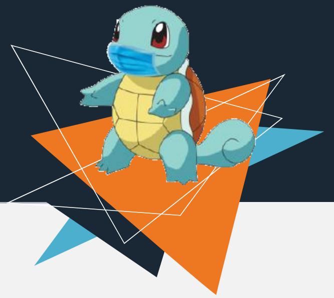


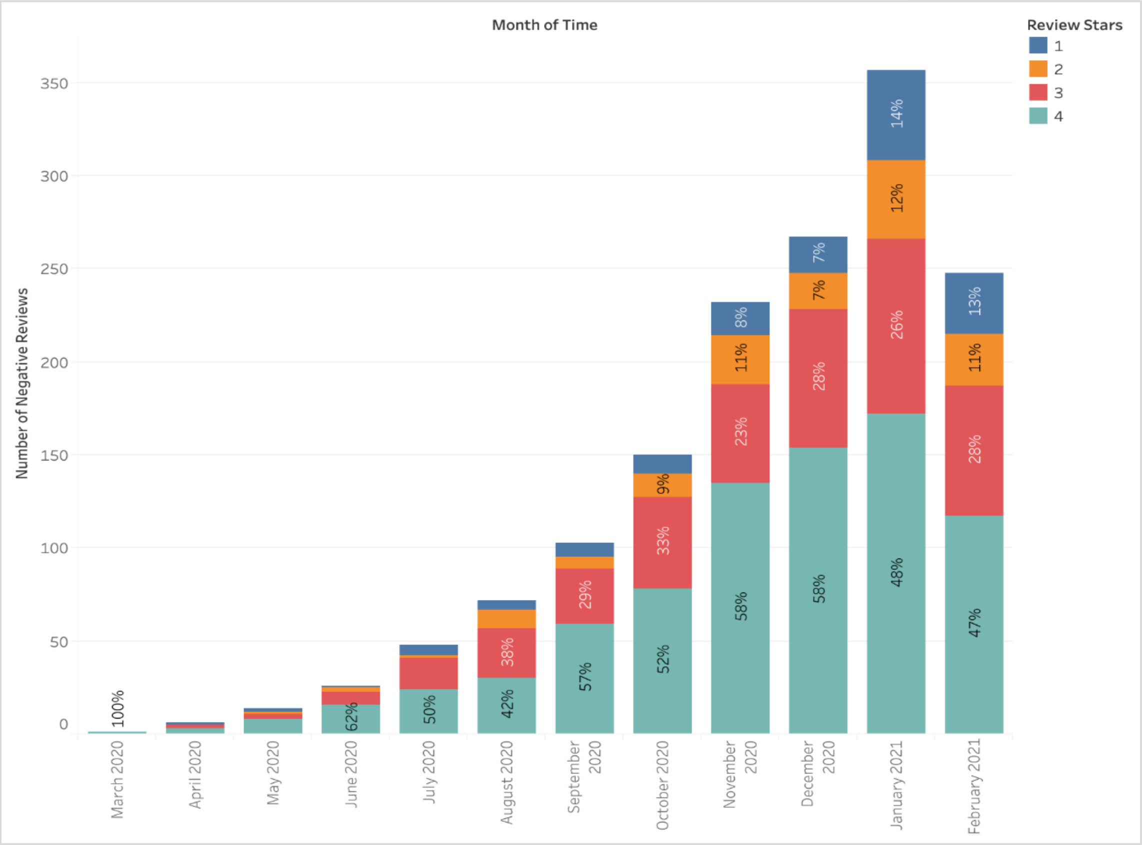
Example High Score Reviews with Negative Sentiment

*"Delivery was ok. **Box came dented** all over but mask was ok n in good condition. **Thin mask.** No funny smell. Elasticity of the loop is good."*

*"Ordered on 6th, rec'd parcel on 10 Feb at 11.30pm. Bot 4 boxes, **1 box came empty.** Not sure why can pack till like that? Mail out empty box. **Msg seller completely no response** at all. Filed for refund, tdy 16 Feb deadline so shopee auto approve refund since no response fr seller. Masks quality ok."*

EDA: Increment on Low-Star Negative Reviews





Negative Reviews along Time

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

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

EDA: Reviews on Multiple Aspects

A negative customer review usually does not focus on only one aspect.

- There is a need to identify the corresponding paint point aspect behind each customer review in order to improve the sellers’ and the platform's performance.

Delivery and Product

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

Delivery, Product & Service

*“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.”*

Delivery and Service

*“ Ordered on 3/12/20, received on 8/12/20. **Why so slow?** Despite slow process from seller, mask seems good, still can smell my fart. Though that's not the point. One of the cheapest, only **drawback is seller response.** Doing this for 7 coins. Nothing else.”*

Product and Service

“ Looks good! Worth the money. Quality seems good too.

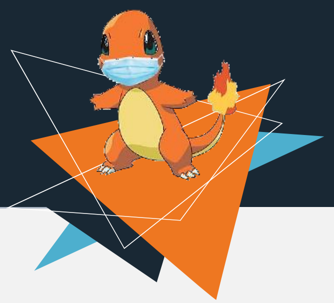
*Edit: realised 20 out of the 30 masks **don’t have the metal part for the nose**. Reached out to them and they **didn’t responded** even though they are online. BUY AT YOUR OWN RISK!”*

03

Sentiment Analysis

To automatically split negative / positive reviews

Problem Description



01

Sentiment Analysis

To separate positive and negative customer reviews automatically.

02

Unsupervised ML

To cluster different negative reviews with respect to their reviewing aspects.

03

Supervised ML

To classify future negative customer reviews into corresponding aspects.

Challenges

- Need to recognize any negative sentiment that exists in high score reviews whose overall sentiment is positive.
- No sentiment target labels. Only rating stars are available.

Solution: Sentiment-based Classification

- Assume that rating star is highly correlated with sentiment, so it can be used to decide the initial sentiment target labels.
- Build a classifier to predict customers' sentiment, so that positive/negative reviews can automatically be separated.
- Only low certainty prediction results need to be manually adjusted, avoiding the need of manual work on assigning sentiment labels to the whole dataset.

Model Development

Training Data

- All 1-3 stars reviews, in total 760 are used.
- 800 of all 5-star reviews are sampled to form a balanced training data.
- 4-star reviews are not used because it is highly mixed of positive and negative comments.

Data Cleaning

- Lemmatization
- **Each customer review is split into sub-sentences**
- Sentences are split with punctuations and words such as 'however', 'but' …

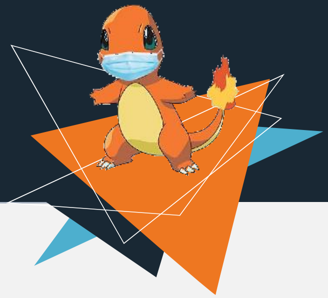
Sentiment Analysis

- *nltk.sentence_tokenize* and *SentimentIntensityAnalyzer* for sentiment analysis
- **Sentiment target labels are assigned for training data**
- For **1-3 star reviews**, positive sub-sentences are removed, then assign **'negative'** label.
- For **5 star reviews**, negative sub-sentences are removed, then assign **'positive'** label.

Model Training

- Bayes optimization
- CountVectorizer, and min_df = 2
- MultinomialNB, alpha = 0.70344
- 5-Fold cross validation

Sentiment Classification Results



Prediction

The trained model is used to predict whether there is any negative sentiment for all 1-4 star reviews.

- 5-star reviews are assumed non-negative so excluded in following studies.
- Manual adjustments on low-certainty predictions are done to get true-negative dataset.
- Evaluations are based on the predicted labels and the true labels.

Scores of Proposed Method

	Precision	Recall	F1-score
0 (negative)	0.96	0.89	0.93
1 (positive)	0.68	0.87	0.76
Accuracy			0.89
Macro avg	0.82	0.88	0.85
Weighted avg	0.90	0.89	0.89

Scores of Solely Sentiment Analysis***

	Precision	Recall	F1-score
0 (negative)	0.97	0.50	0.66
1 (positive)	0.41	0.96	0.58
Accuracy			0.62
Macro avg	0.69	0.73	0.62
Weighted avg	0.83	0.62	0.64

****For comparison purpose only. These results are based on solely sentiment analysis on the whole sentences of the customer reviews.*

Conclusions

High accuracy with a small amount of training data, can be efficient for large datasets.

Outperforms pure sentiment analysis, enabling less manual work on separating negative and positive reviews.

Word Cloud of All Negative Reviews

Mainly 3 aspects are found.

Delivery
- Slow / wrong delivery...
- Dented box...
- …

Product
- Smell / itchy nose/ thin...
- Price too high...
- …

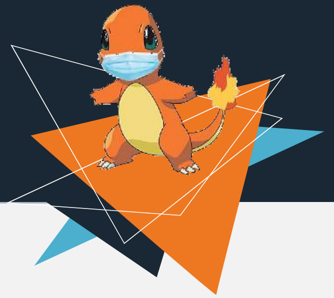
Service
- Wrong color sent / badly packed...
- No response / no refund...
- ...

04

Unsupervised ML

Aspect-based clustering of negative reviews

Problem Description



01

Sentiment Analysis

To separate positive and negative customer reviews automatically.

02

Unsupervised ML

To cluster different negative reviews with respect to their reviewing aspects.

03

Supervised ML

To classify future negative customer reviews into corresponding aspects.

Challenges

- The text clustering task is always challenging because of the word sequence, word context, mis-spelling…
- KMeans tends to get clusters with similar amount of data, but data overview shows that the numbers of negative reviews related to different aspects are quite different.

Solution: Similarity-based Clustering

- Determine initial centroids with high confidence data
- Similarity-based clustering for the rest data
- Finally, manual adjustment for 'blurred' customer reviews whose similarities to more than one clusters are close.

Model Development

01

BERT Vector

- Use only the negative part of each customer reviews.
- Use 1st last layer of Bert Model to get output vectors.

02

Initial Centroids

- k=3 to include delivery, product and service
- Aspect-based keywords to find initial 'core data' for each cluster.
- Find centroid of each cluster by taking average of the BERT vector of the 'core data'.

03

Similarity Analysis

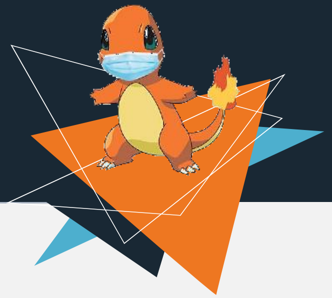
- Compute cosine similarity of each negative review to each of the cluster centroids.
- Assign each review to its closest cluster.

04

Improvements

- Further improvements are possible, including the following.
- Keep update aspect-based keyword lists.
- Iteratively update the cluster centroids as more data are added to each of the clusters

Aspect-based Clustering Results



- Manual adjustment to get true aspect labels for 'blurred' customer reviews after the clustering analysis.
- Evaluations are based on the clustered labels and the true labels.

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

** Kmeans clustering case for comparison purpose.*

*** Base case of the proposed method uses only one keyword to determine the initial 'core data' for each cluster, but can achieve relatively high accuracy results.*
delivery_words = ['delivery’] ; product_words = ['quality’]; service_words = ['service']

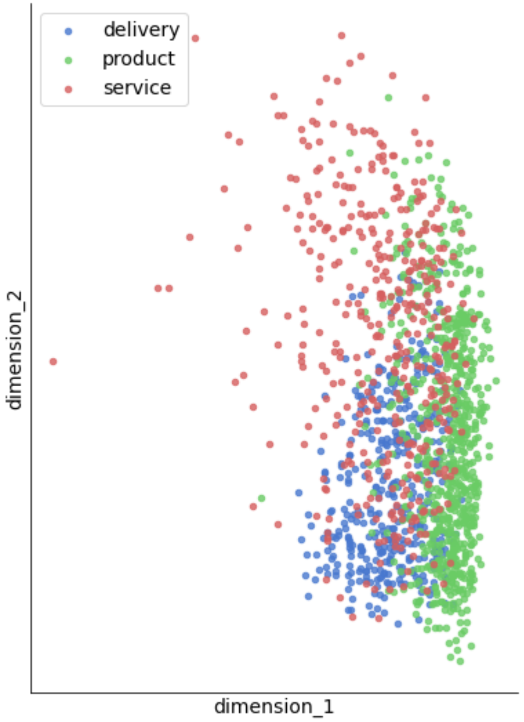
*** Improved case uses more keywords to achieve more accurate results.*
delivery_words = ['box', 'date', 'days', 'delivery', 'dented', 'received', 'plastic', 'sealed', 'time', 'week']
product_words = ['2nd', 'big', 'black', 'different', 'easily', 'loose’, 'material', 'nose', 'product', 'quality', 'size', 'small', 'smell', 'soft', 'surgical', 'thick', 'thin', 'tight']
service_words = ['service']

Conclusions

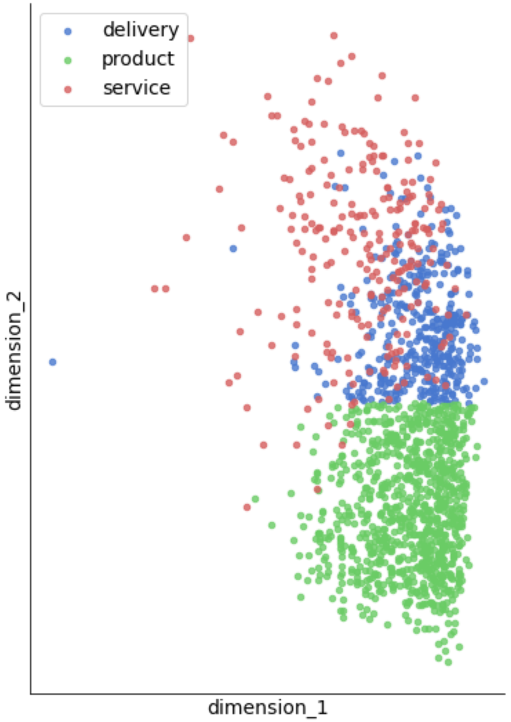
- The proposed method outperforms conventional KMeans method.
- Acceptable accuracy even with only a single word as a start in the keyword list, efficient for large datasets.

Aspect-based Clustering Visualization

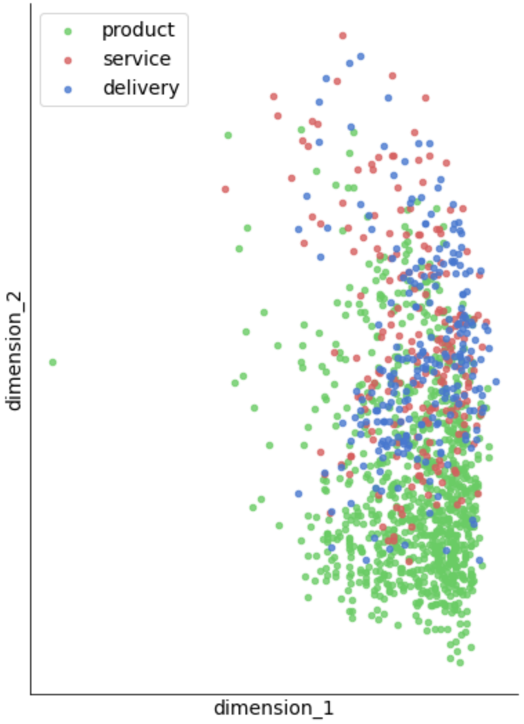
The proposed method obviously outperforms the KMeans.



KMeans



Improved Case



True Labels*

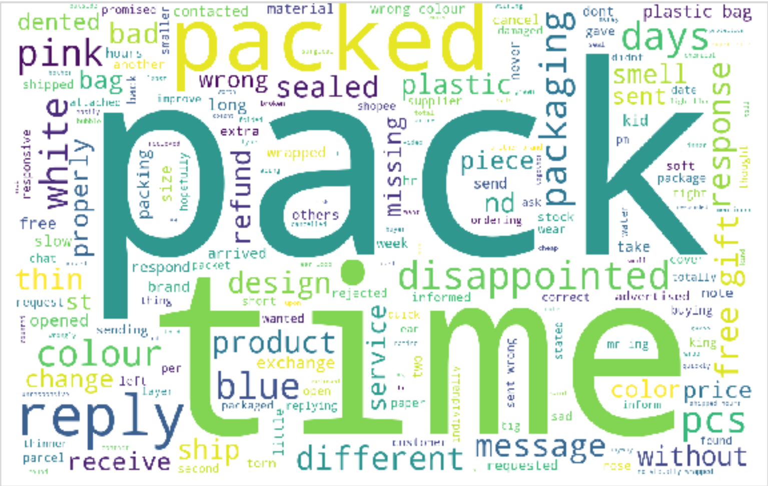
Notes

- The 2 dimensions are transformed values based on similarities to the cluster centroids.
- *True labels are actually multi-aspects while predicted labels are single aspect. Those minor multi-aspect data are excluded for comparison purpose.

Aspect-based Word Clouds

Delivery

Service



Product

Comparison

05

Supervised ML

Aspect-based prediction of negative reviews

Problem Description

01

Sentiment Analysis

To separate positive and negative customer reviews automatically.

02

Unsupervised ML

To cluster different negative reviews with respect to their reviewing aspects.

03

Supervised ML

To classify future negative customer reviews into corresponding aspects.

Review Text Data

Bert Encoder

Word Vectors from Bert

Features

Classifiers

Delivery | Product | Service

(multiple labels)

Naïve Bayes

KNN

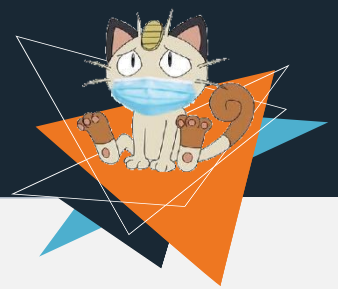
BERT

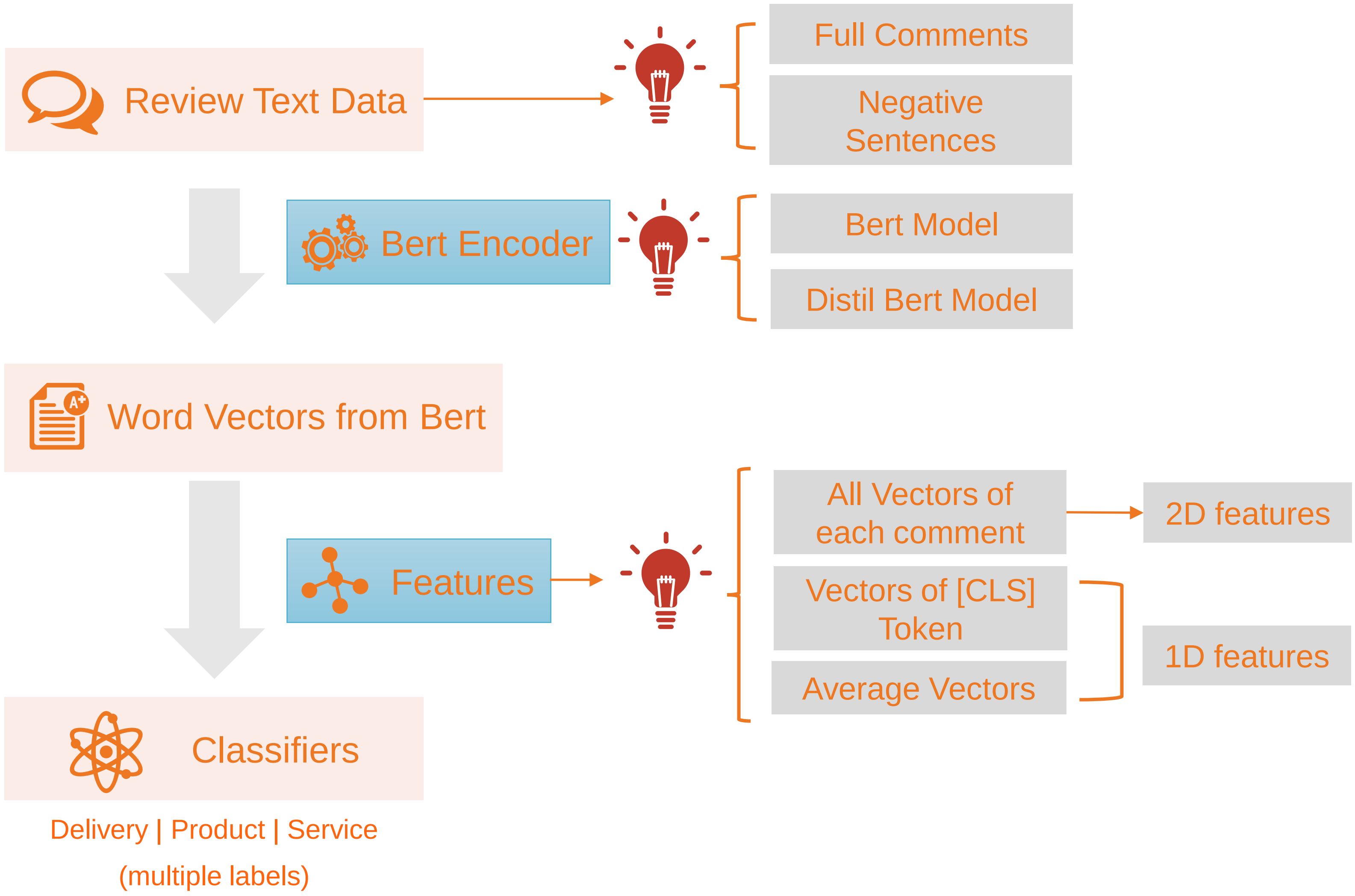
CNN

RNN

LSTM

Methodology





Benchmark Models

- NB
- KNN

Multiple cases

- 1D ([CLS], avg, max) / 2D
- Full / Negative Comments
- BertModel / DistilBert

LSTM to select the best case

Deep Learning Models to run the best case

- LSTM
- BERT
- CNN
- RNN

Parameter Study Results (Selected Cases)

- **Baseline Model: LSTM**

Options	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
Feature Dimension	1D ([CLS])	1D (avg)	2D	1D ([CLS])	1D (avg)	2D
Sentence	Full	Full	Full	Negative	Negative	Negative
Encoder	DistilBert	DistilBert	DistilBert	DistilBert	DistilBert	DistilBert
Macro f1	0.79|0.82|0.779	0.805|0.851|0.784	0.813|0.83|00.747	0.804|0.83|00.723	0.838|0.872|0.761	0.846|0.836|0.749
AUC	0.899|0.909|0.846	0.908|0.925|0.877	0.915|0.924|0.867	0.896|0.916|0.846	0.932|0.929|0.844	0.931|0.923|0.831

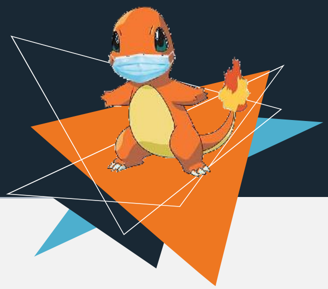
Model Selection Results (Selected Cases)

BERT Classifier outperforms the others

- **Multi-head Attention mechanism** of BERT layer in the encoder can better capture semantic meanings and similarity
- Model weights in BERT layers are **fine-tuned** with new data based on the pre-trained model

	NB	KNN	LSTM	CNN	RNN	BERT Classifier
Feature Dimension	1D (avg)	1D (avg)	1D (avg)	2D	2D	/
Sentence	Negative	Negative	Negative	Negative	Negative	**Negative**
Encoder	DistilBert	DistilBert	DistilBert	DistilBert	DistilBert	Fine-tuned Bert Layers
Macro f1	0.735|0.781|0.682	0.73|0.808|0.692	0.838|0.872|0.761	0.873|0.877|0.773	0.802|0.854|0.741	**0.873|0.886|0.803**
AUC	0.843|0.878|0.797	0.866|0.911|0.832	0.932|0.929|0.844	0.933|0.933|0.869	0.897|0.911|0.816	**0.943|0.941|0.850**

Further Discussions



Text Classification is challenging

- Some cases are difficult to be classified even manually.
- Therefore, sometimes misclassification is understandable, and there is room for further improvements.

Example 01

Complex context

'The glue smell the moment i open the box. Dont understand the label in pic 1. 19083 is a surgical mask standard, yet implement is GB/T32610 which is a non medical mask? Mask look genuine, but the natnl standard is confusing.'

Example 02

Ambiguous attitude

Received the Masks today but it is different from what i ordered 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

Example 03

Fuzzy aspect

'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 with double sided type and the box is flattened..'

Conclusions

- The current models can achieve acceptable performance despite of some tolerable errors.
- Final true labels can be obtained by manual adjustment of 'blurred' probability labels.
- Possible improvements:
  training with a larger dataset to identify more detailed aspects of complaints;
  adding manual features to help the model identify context-specific expressions.

06

Extended Studies

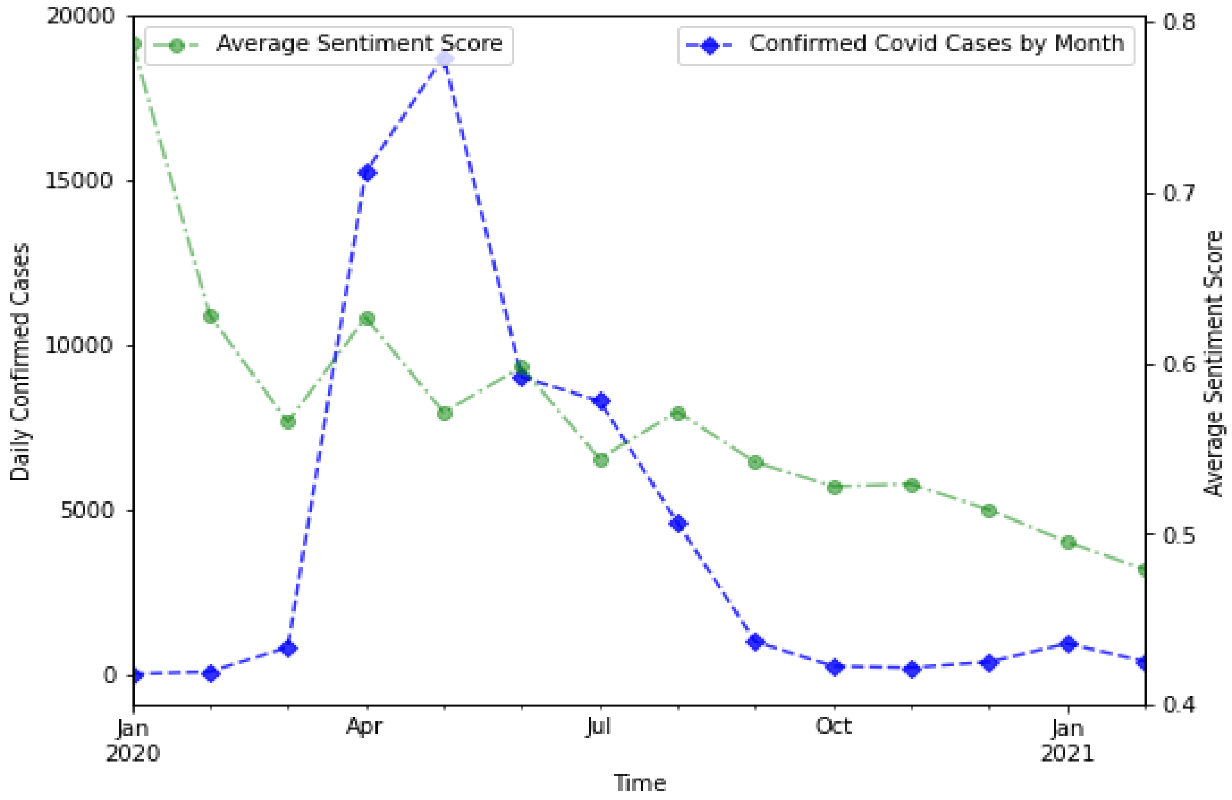
To extract business insights

Covid Cases & Average Sentiment Score

There is a decreasing trend in the customers' sentiment over time in general.

- But note that there is a lag between a customer's purchasing time and reviewing time.
- 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, customers are still demanding.

Covid Cases & Average Sentiment Score* along Time



*Average sentiment score by month is calculated from all reviews (1-5 stars) to reflect the overall sentiment of the customers.

Aspect-based Average Sentiment Scores

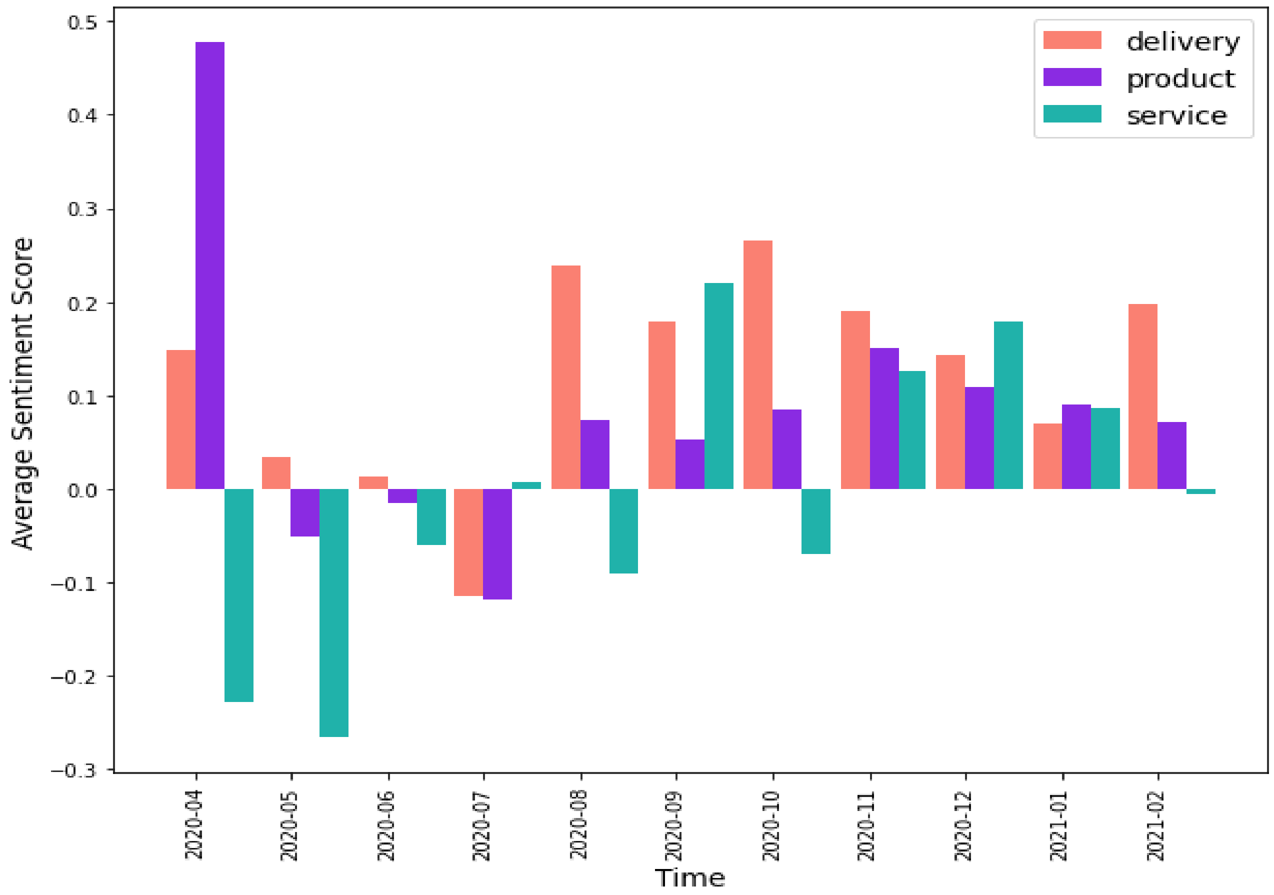
Sentiments scores are very low among all aspects during the peak time of covid cases.

- Aspect labels are available for **negative reviews** based on the outcome in the previous analysis sections.

- The average sentiment score by aspect is calculated from **negative reviews** w.r.t each aspect.

- As the covid severeness has been mitigated, the scores for all three aspects gradually improve.

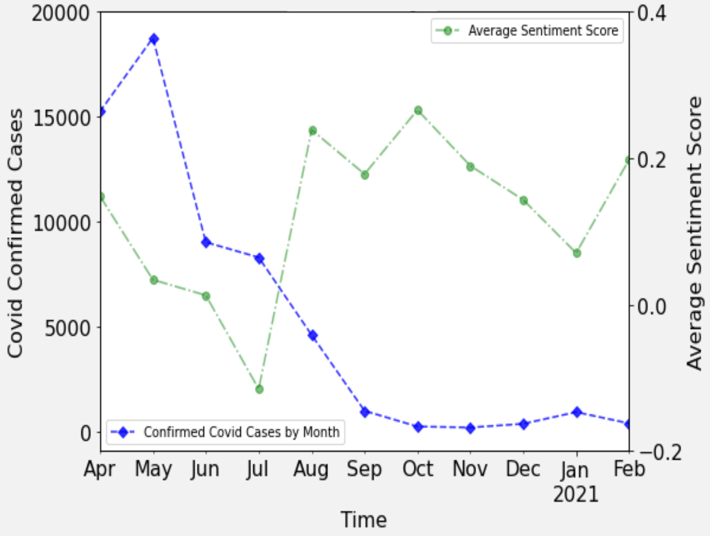
Problem solving priority can be based on the negative scores in different periods.



Covid Cases & Sentiment Score by Aspects

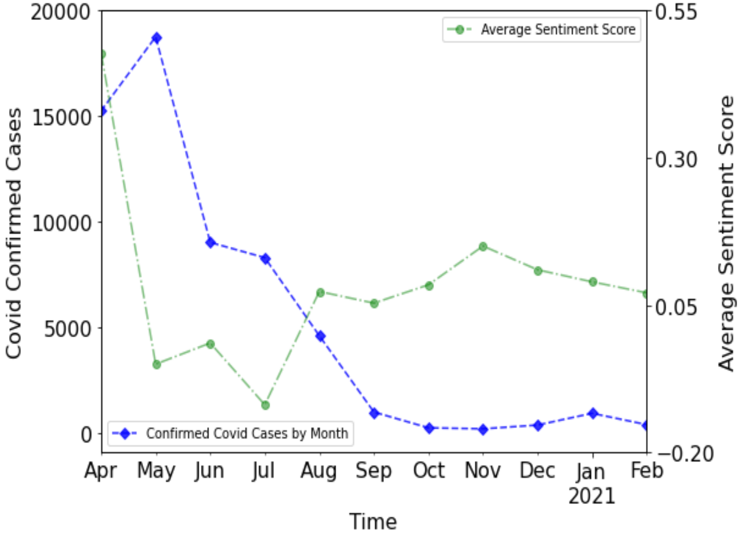
Delivery

- Initially, the sentiment score drops as the number of covid cases remain high.
- Possibly when the government enforced mask-wearing, customers hoped to receive the masks fast (more demanding on delivery).
- The score improves as covid cases fall.



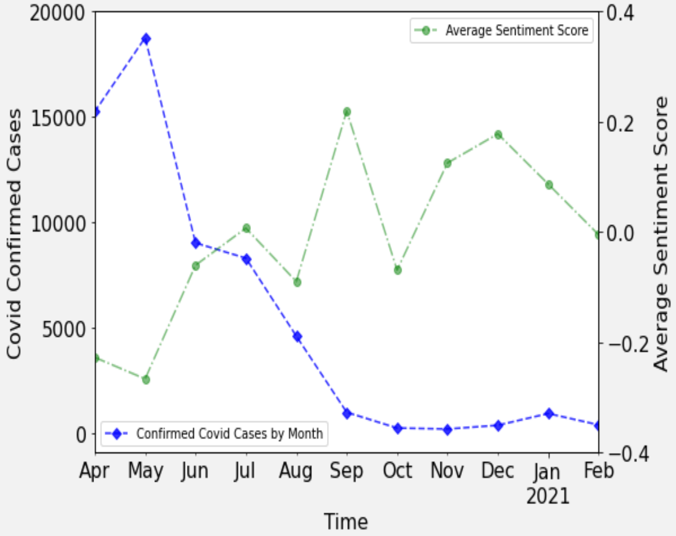
Product

- A huge drop when covid cases reached the peak.
- The score is still relatively low (~0.05) even when the number of cases decreased, which may indicate that customers remain rigorous on the product aspect (eg. Quality).



Service

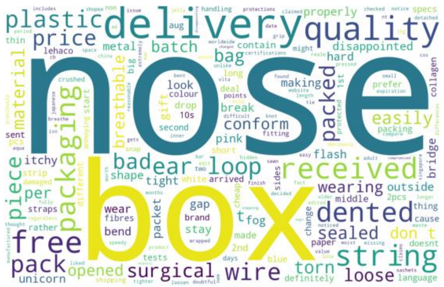
- In general shows an increasing trend.
- Possibly because the platform and the sellers improved their services according to customers' negative reviews.



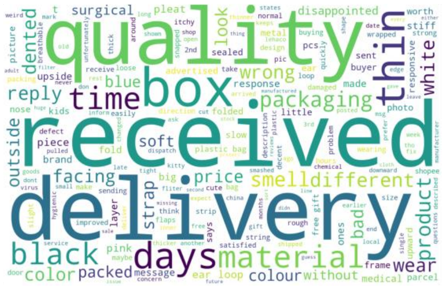
Word Cloud of Negative Reviews by Month

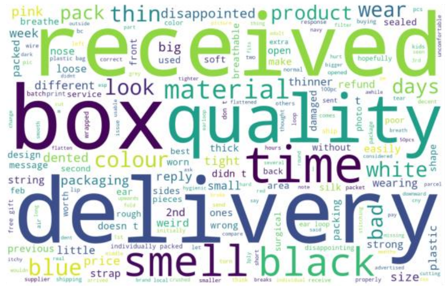
- Customers focused more on **delivery & packing** during the early stage of the pandemic
- Gradually placed more emphasis on mask **quality** at the later stage.
- The outcome of the word clouds is aligned with the previous discussions.

Apr 2020



Sept 2020





Feb 2020

07

Summary

Highlights

Methodology Framework

- A 3-stage methodology framework is proposed to identify the pain points behind the customers' negative reviews.

Business Insight

- The outcome in this study helps to target specific problems and improve overall customer services accordingly.

Generalization Ability

- The analytic approach can be generalized to larger datasets and to other industries easily.

Possible Improvements

01 Larger Datasets

- The current dataset is relatively small because negative reviews are rare in all reviews.
- Can include review from multiple platforms or different types of products.

02 Methodology Improvements

- Improve unsupervised ML by iteratively updating each cluster's centroid as data are added in.
- Improve supervised ML by adding manual features to help identify context-specific representations.

03 Model Optimization

- More comprehensive parameter optimization for the models.
- More state-of-the-art models to be included to improve the prediction performances.

THANK YOU

Source code on Github: https://github.com/justdropby/BT5153_Group11