Reviews

A User Generated Issue

Ken Tan
DSI-25

Table of Contents

01 Introduction

02 Data Cleaning & Exploratory Data Analysis

03 Modelling

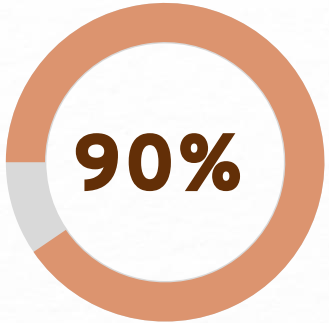
04 Conclusion, Recommendations & Future Outlook

01

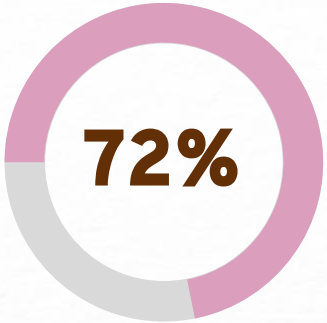
Introduction

Product Reviews

Customer opinions or feedback for a particular product



Read online reviews



Take action only after reading a positive review

https://www.invespcro.com/blog/the-importance-of-online-customer-reviews-infographic/

Shopee Product Reviews?

y*****y

★★★★★

fast delivery, item arrived on the day it was shipped out!!!

x*****7

★★★★★

Havent tried to cook using the item but hopefully is good. Thank you.

g*****y

★★★★★

Variation: 15cm | 500g,5 Aug

Yet to try as just received! Customer service is great and responsive. Requested for no knife as I've lots at home.

Problem Statement

Product reviews are an essential part of any business. Moreso, for e-commerce sites since pictures and descriptions provided by the sellers may not be an accurate representation of a product. As there are no other ways for consumers to verify or try products out, reviews become an important supplement source of information.

As a 3rd party analyst on business improvements, we aim to detect real and fake reviews using Natural Language Processing (NLP) by analysing data from pulled from Shopee's site.

02

Data Cleaning & Exploratory Data Analysis

Label

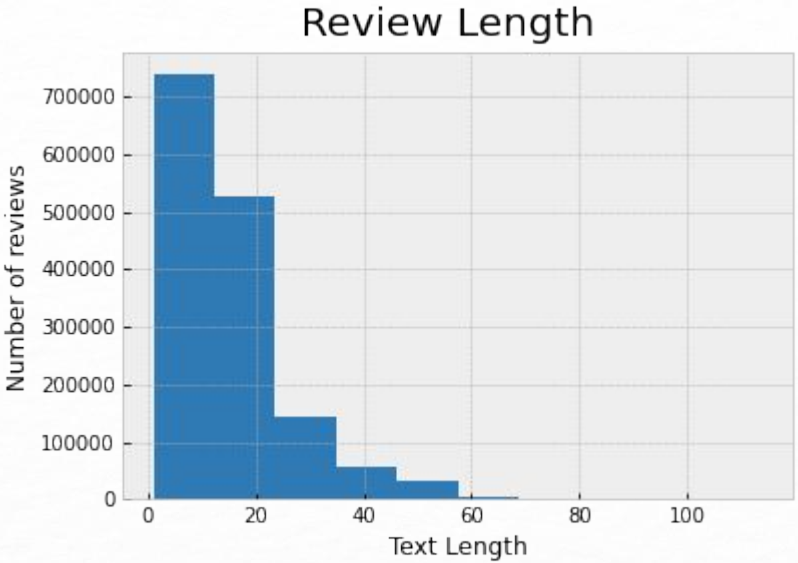
Text

Preliminary Look

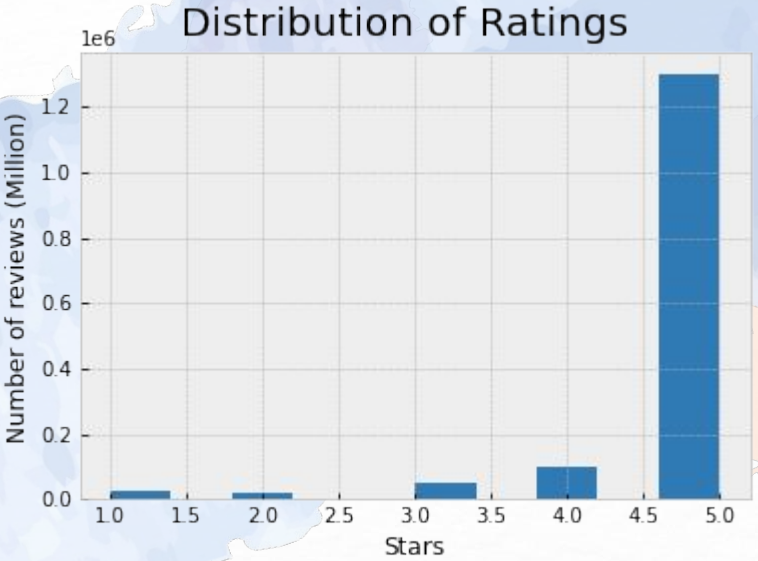
	label	text
0	5	Looks ok. Not like so durable. Will hv to use a while to recommend others of its worth.
1	5	Tried, the current can be very powerful depending on the setting, i don't dare to go higher but if go higher sure muscle will become sore and can see the effect faster.
2	5	Item received after a week. Looks smaller than expected, can't wait to try!
3	5	Thanks!!! Works as describe no complaints. Not really expecting any life changing results but thanks!
4	5	Fast delivery considering it's from overseas and only tried once. Not sure about the results yet.
5	5	Fast delivery good service
6	5	Got my order and it came well packaged. Have yet to try but looks good so far. Thanks!
7	5	Items received in a nice box. Have not used it yet, hopefully it works!
8	5	Received in good condition, tried so far so good. Not that bad.
9	1	Item doesn't work . \n\nAsked me to send a refund , show a non working machine and deem not enough evidence . \n\nDon't waste time buying .

Preliminary Data Cleaning

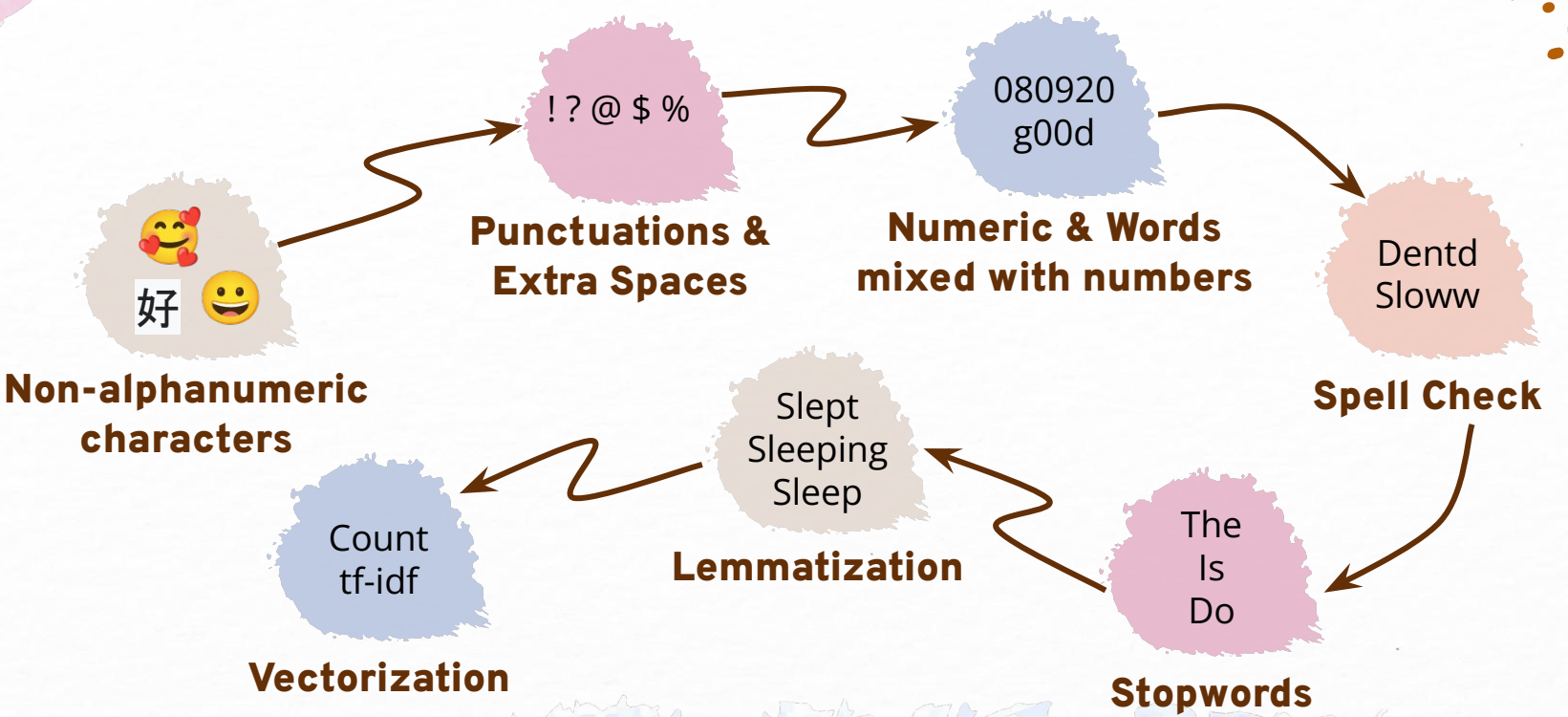
Exploratory Data Analysis



Exploratory Data Analysis



Textual Cleaning & Preprocessing



Top Words (Unigram)

1. Good

Top Words (Unigram)

1. Good
2. Delivery

Top Words (Unigram)

1. Good
2. Delivery
3. Received

Top Words (Unigram)

1. Good
2. Delivery
3. Received
4. Item

Top Words (Unigram)

1. Good
2. Delivery
3. Received
4. Item
5. Fast

Top Words (Bigram)

1. Fast Delivery

Top Words (Bigram)

1. Fast Delivery
2. Good Condition

Top Words (Bigram)

1. Fast Delivery
2. Good Condition
3. Received Good

Top Words (Bigram)

1. Fast Delivery
2. Good Condition
3. Received Good
4. Item Received

Top Words (Bigram)

1. Fast Delivery
2. Good Condition
3. Received Good
4. Item Received
5. Good Quality

03

Modelling

Manual Labelling

Sample Size

1000

text	text_length	real_review
delivery time was decent the scrub smells good and skin doesnt feel dry after using	15	1
came as described well packed with a little spatula great	10	1
fast delivery good size for babies quality is good affordable	10	1
received with thanks	12	0
very prompt delivery and response items neatly and hygienically packed nappropriate zipbloc my first time ordering	16	0
thank you very much fast shipping as well received in good condition	12	0

Modelling with Pycaret!

CountVectorizer unigram

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lr	Logistic Regression	0.7996	0.8497	0.8807	0.8276	0.8529	0.5385	0.5430	0.0410
xgboost	Extreme Gradient Boosting	0.7925	0.8673	0.8657	0.8291	0.8461	0.5266	0.5310	0.5340
catboost	CatBoost Classifier	0.7868	0.8450	0.9436	0.7801	0.8539	0.4717	0.5043	1.2420
et	Extra Trees Classifier	0.7767	0.8413	0.8309	0.8309	0.8299	0.5038	0.5060	0.1210
rf	Random Forest Classifier	0.7725	0.8331	0.8395	0.8208	0.8289	0.4874	0.4906	0.1140
gbc	Gradient Boosting Classifier	0.7654	0.8242	0.9285	0.7669	0.8397	0.4160	0.4451	0.1560
ada	Ada Boost Classifier	0.7653	0.8206	0.8376	0.8141	0.8247	0.4684	0.4715	0.0700
svm	SVM - Linear Kernel	0.7610	0.0000	0.8071	0.8291	0.8165	0.4718	0.4753	0.0270
lightgbm	Light Gradient Boosting Machine	0.7511	0.8104	0.8572	0.7864	0.8199	0.4186	0.4239	0.1300
ridge	Ridge Classifier	0.7352	0.0000	0.7940	0.8022	0.7975	0.4143	0.4156	0.0270
dt	Decision Tree Classifier	0.7096	0.6825	0.7660	0.7901	0.7761	0.3606	0.3644	0.0270
knn	K Neighbors Classifier	0.7009	0.7381	0.7531	0.7843	0.7674	0.3476	0.3489	0.0340
nb	Naive Bayes	0.5807	0.5848	0.5691	0.7399	0.6412	0.1539	0.1639	0.0260
lda	Linear Discriminant Analysis	0.5522	0.5427	0.5865	0.6915	0.6333	0.0644	0.0669	0.1070
qda	Quadratic Discriminant Analysis	0.3690	0.5165	0.0581	0.3727	0.0958	0.0241	0.0412	0.0600

tf-idf Vectorizer unigram

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
et	Extra Trees Classifier	0.7910	0.8108	0.9091	0.8031	0.8522	0.4987	0.5130	0.1240
ridge	Ridge Classifier	0.7896	0.0000	0.9220	0.7958	0.8533	0.4868	0.5096	0.0260
lr	Logistic Regression	0.7825	0.8591	0.9675	0.7667	0.8550	0.4406	0.4971	0.6800
xgboost	Extreme Gradient Boosting	0.7725	0.8353	0.8679	0.8052	0.8338	0.4712	0.4814	0.6170
catboost	CatBoost Classifier	0.7695	0.8379	0.9110	0.7793	0.8388	0.4385	0.4668	2.0410
svm	SVM - Linear Kernel	0.7667	0.0000	0.8398	0.8153	0.8253	0.4708	0.4795	0.0290
rf	Random Forest Classifier	0.7653	0.8049	0.9003	0.7798	0.8352	0.4325	0.4515	0.1010
ada	Ada Boost Classifier	0.7637	0.8079	0.8461	0.8079	0.8250	0.4591	0.4660	0.0720
lightgbm	Light Gradient Boosting Machine	0.7610	0.8113	0.8595	0.7965	0.8259	0.4443	0.4508	0.2040
gbc	Gradient Boosting Classifier	0.7553	0.8066	0.8960	0.7712	0.8286	0.4078	0.4262	0.1800
dt	Decision Tree Classifier	0.7166	0.6771	0.7992	0.7802	0.7887	0.3575	0.3597	0.0310
knn	K Neighbors Classifier	0.6207	0.6601	0.6347	0.7576	0.6814	0.2140	0.2242	0.3790
nb	Naive Bayes	0.6022	0.6027	0.6047	0.7488	0.6682	0.1843	0.1926	0.0250
lda	Linear Discriminant Analysis	0.6022	0.5884	0.6457	0.7237	0.6810	0.1555	0.1578	0.1290
qda	Quadratic Discriminant Analysis	0.4076	0.5273	0.1598	0.7627	0.2386	0.0417	0.0725	0.0620

CountVectorizer Bigram

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lr	Logistic Regression	0.7382	0.7597	0.9352	0.7397	0.8257	0.3279	0.3691	0.7650
ada	Ada Boost Classifier	0.7182	0.6935	0.9158	0.7295	0.8112	0.2802	0.3165	0.1000
et	Extra Trees Classifier	0.7182	0.6924	0.8316	0.7654	0.7959	0.3404	0.3465	0.1640
gbc	Gradient Boosting Classifier	0.7139	0.6967	0.9460	0.7152	0.8142	0.2404	0.2945	0.1900
catboost	CatBoost Classifier	0.7095	0.7374	0.9568	0.7082	0.8136	0.2164	0.2758	0.9220
ridge	Ridge Classifier	0.7081	0.0000	0.8596	0.7419	0.7962	0.2918	0.3016	0.0310
rf	Random Forest Classifier	0.7067	0.7106	0.8576	0.7425	0.7952	0.2866	0.2959	0.1380
svm	SVM - Linear Kernel	0.7038	0.0000	0.8142	0.7574	0.7840	0.3131	0.3169	0.0340
xgboost	Extreme Gradient Boosting	0.6952	0.6951	0.8901	0.7183	0.7941	0.2301	0.2567	1.1080
dt	Decision Tree Classifier	0.6867	0.6417	0.8167	0.7391	0.7748	0.2620	0.2692	0.0430
knn	K Neighbors Classifier	0.6652	0.5822	0.9099	0.6858	0.7807	0.1169	0.1562	0.3860
lightgbm	Light Gradient Boosting Machine	0.6624	0.5988	0.9461	0.6754	0.7877	0.0617	0.0855	0.2010
lda	Linear Discriminant Analysis	0.6409	0.6271	0.7667	0.7129	0.7367	0.1680	0.1733	0.1570
nb	Naive Bayes	0.5751	0.6117	0.4988	0.7825	0.6068	0.1904	0.2153	0.0290
qda	Quadratic Discriminant Analysis	0.3676	0.4482	0.1985	0.4742	0.2555	-0.0789	-0.1416	0.0950

tf-idf Vectorizer bigram

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
ridge	Ridge Classifier	0.7439	0.0000	0.9634	0.7223	0.8251	0.3810	0.4456	0.0300
rf	Random Forest Classifier	0.7425	0.7700	0.8860	0.7502	0.8116	0.4124	0.4305	0.1520
et	Extra Trees Classifier	0.7395	0.7673	0.8631	0.7583	0.8057	0.4137	0.4262	0.2020
svm	SVM - Linear Kernel	0.7339	0.0000	0.9132	0.7333	0.8113	0.3746	0.4090	0.0430
catboost	CatBoost Classifier	0.7339	0.7786	0.9406	0.7208	0.8159	0.3648	0.4119	1.3510
gbc	Gradient Boosting Classifier	0.7310	0.7525	0.9270	0.7232	0.8119	0.3637	0.4057	0.2050
ada	Ada Boost Classifier	0.7296	0.7176	0.8997	0.7324	0.8066	0.3714	0.3962	0.1080
dt	Decision Tree Classifier	0.7138	0.6804	0.8425	0.7399	0.7868	0.3559	0.3651	0.0450
lr	Logistic Regression	0.7053	0.7808	0.9886	0.6837	0.8081	0.2553	0.3568	0.0530
xgboost	Extreme Gradient Boosting	0.6924	0.7107	0.8540	0.7133	0.7767	0.2940	0.3083	1.1620
lda	Linear Discriminant Analysis	0.6681	0.6463	0.7832	0.7164	0.7466	0.2665	0.2723	0.1670
lightgbm	Light Gradient Boosting Machine	0.6610	0.6292	0.9041	0.6704	0.7694	0.1777	0.2135	0.1530
knn	K Neighbors Classifier	0.6566	0.6164	0.9477	0.6571	0.7757	0.1362	0.1939	0.0480
nb	Naive Bayes	0.5807	0.6098	0.4977	0.7511	0.5972	0.1956	0.2151	0.0300
qda	Quadratic Discriminant Analysis	0.3777	0.4724	0.0983	0.5050	0.1640	-0.0428	-0.0864	0.1070

Model Comparison

CountVectorizer unigram

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lr	Logistic Regression	0.7996	0.8497	0.8807	0.8276	0.8529	0.5385	0.5430	0.0410
xgboost	Extreme Gradient Boosting	0.7925	0.8673	0.8657	0.8291	0.8461	0.5266	0.5310	0.5340
catboost	CatBoost Classifier	0.7868	0.8450	0.9436	0.7801	0.8539	0.4717	0.5043	1.2420

Tf-idf Vectorizer unigram

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
et	Extra Trees Classifier	0.7910	0.8108	0.9091	0.8031	0.8522	0.4987	0.5130	0.1240
ridge	Ridge Classifier	0.7896	0.0000	0.9220	0.7958	0.8533	0.4868	0.5096	0.0260
lr	Logistic Regression	0.7825	0.8591	0.9675	0.7667	0.8550	0.4406	0.4971	0.6800

Tf-idf Vectorizer bigram

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
ridge	Ridge Classifier	0.7439	0.0000	0.9634	0.7223	0.8251	0.3810	0.4456	0.0300
rf	Random Forest Classifier	0.7425	0.7700	0.8860	0.7502	0.8116	0.4124	0.4305	0.1520
et	Extra Trees Classifier	0.7395	0.7673	0.8631	0.7583	0.8057	0.4137	0.4262	0.2020

CountVectorizer Bigram

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lr	Logistic Regression	0.7382	0.7597	0.9352	0.7397	0.8257	0.3279	0.3691	0.7650
ada	Ada Boost Classifier	0.7182	0.6935	0.9158	0.7295	0.8112	0.2802	0.3165	0.1000
et	Extra Trees Classifier	0.7182	0.6924	0.8316	0.7654	0.7959	0.3404	0.3465	0.1640

Model Tuning

Hyperparameter Tuning

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.8143	0.8877	0.9130	0.8235	0.8660	0.5663	0.5743
1	0.8571	0.8850	0.9565	0.8462	0.8980	0.6628	0.6768
2	0.8143	0.8986	0.9783	0.7895	0.8738	0.5371	0.5838
3	0.8429	0.8197	0.9565	0.8302	0.8889	0.6251	0.6437
4	0.7714	0.8804	0.9348	0.7679	0.8431	0.4366	0.4665
5	0.7857	0.8524	0.9130	0.7925	0.8485	0.4888	0.5033
6	0.7714	0.8514	0.9348	0.7679	0.8431	0.4366	0.4665
7	0.7714	0.8143	0.8696	0.8000	0.8333	0.4717	0.4759
8	0.8429	0.8813	0.9348	0.8431	0.8866	0.6330	0.6420
9	0.8261	0.8546	0.9111	0.8367	0.8723	0.6012	0.6065
Mean	0.8098	0.8625	0.9302	0.8097	0.8654	0.5459	0.5639
SD	0.0312	0.0275	0.0290	0.0284	0.0212	0.0798	0.0761

LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, l1_ratio=None, max_iter=1000, multi_class='auto', n_jobs=None, penalty='l2', random_state=154692427, solver='lbfgs', tol=0.0001, verbose=0, warm_start=False)

Model Prediction

```
predictions = predict_model(lr_final)
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Logistic Regression	0.9236	0.9732	0.9695	0.9183	0.9432	0.8267	0.8296

04

Conclusion
Recommendations
Future Outlook

Conclusion

	First Quarter	Second Quarter	Third Quarter
Key Action 1	Here you can describe your items for the quarter	Here you can describe your items for the quarter	Here you can describe your items for the quarter
Key Action 2	Here you can describe your items for the quarter	Here you can describe your items for the quarter	Here you can describe your items for the quarter
Key Action 3	Here you can describe your items for the quarter	Here you can describe your items for the quarter	Here you can describe your items for the quarter

Recommendations

Tighten Incentives Requirements

Deploy model to only reward incentives to legitimate reviews

Include Other Scoring Metrics for Reviews

Delivery speed and customer service do matter

Increase Time Given for Reviews

Some products like skincare take a while for results to show

Future Outlook

Deep Learning

Use of Neural Networks could possibly help improve predictions

Textual Sentiment Analysis

Despite the large amount of data available, only a small sample was used due to the need for manual labelling

Model Deployment to Lazada

Saturn is the sixth planet from the Sun and the second-largest in the Solar System

More Training Data

Despite the large amount of data available, only a small sample was used due to the need for manual labelling

End

THANKS