UNIVERSITY OF EDINBURGH
Business School

2020-21

Advanced Data Modelling (CMSE11419)

Individual Report Apr. 21

Y. Mahir Nartok

Table of Contents

1. INTRODUCTION

Nowadays, e-commerce has become an indispensable part of the global retail sector (Coppola, 2021). Especially with the pandemic, customer shopping habit has been unreversible changed from traditional and on-shopping to online (PYMNTS, 2020). Thus, customers deliver their orders on their mobile by using firms' applications that they are downloaded by online stores such as Google Play Store and App Store depending on the platform. These online stores with ratings and reviews of customers add value to both the firm/developer and potential new users by providing a crowd-sourced indicator of app quality and customer experiences (Vasa et al., 2012).

As an Amazon e-commerce business analyst, using Google Play Store reviews in the UK, customer feedbacks are analysed, and significant quality issues are presented by using natural language processing tools. Additionally, results are compared with other top competitors running in the same online shopping industry namely eBay, and Argos (Redbytes, 2021). Thus, major problems that customer encounter when using our shop or the features that customers are really like to use are enlightened. Finally, results interpreted to give a managerial insight. Considering the most frequent patterns interpreted by Sentiment Analysis, Topic Search method, and text classification models that customer use in their comment for complaining or praising, these patterns will be our focus in the near future.

2. METHODOLOGY

When collecting data, to avoid bias in the model construction part, a balanced dataset that includes a nearly equal number of reviews from each score are used (7500 total: 500 reviews from each five score for three different companies). For Amazon and eBay, because of high volume, only 2021 reviews are available up until 01 April 2021, but for Argos, the dataset also includes older reviews such as written in 2014.

Reviews are categorised as

- Reviews with scores 1 and 2 as "Low/Negative"
- Reviews with scores 3 and 4 as "Average/Neutral"
- Reviews with a score 5 as "Excellent/Positive" (Mostly focuses on bad reviews so this category is decided smaller than others)

It is started to work on the content column that contains real customers' reviews. Related processes are summarised in *Figure 1* below, and some important steps are explained in detail.

Text normalization and pre-processing stage

- Eliminate everything that does not contribute to explaining customers' feeling
- Remove Stopwords using NLTK English stop word list. For sentiment analysis, words such as 'not' and 'after' are not eliminated.
- Lemmatisation: For topic modelling, just Noun and Adjective are filtered.

Feature Engineering on the cleaned dataset:

- One-Hot Encoding & TF-IDF: the vector design is coded to allow Unigram/Bigram/Trigram will be based for future machine learning models.

- Topics in the reviews are exposed with the help of the LDA method (Bansal, 2016). For Topic Search, to find the optimal number of topics, Intrinsic Measure (UMass) is used to calculated model coherence (Kumar,2018). Only Nouns and Adjectives are considered.
- Polarity scores of each comment are calculated by using Vader and WordNet packages.

Model Constructions:

- With pipeline and cross-validation, with checking supervised classification models' metrics such as accuracy, f-scores, the best model that works on the data will be selected (Bengfort, Bilbro, and Ojeda, 2018).

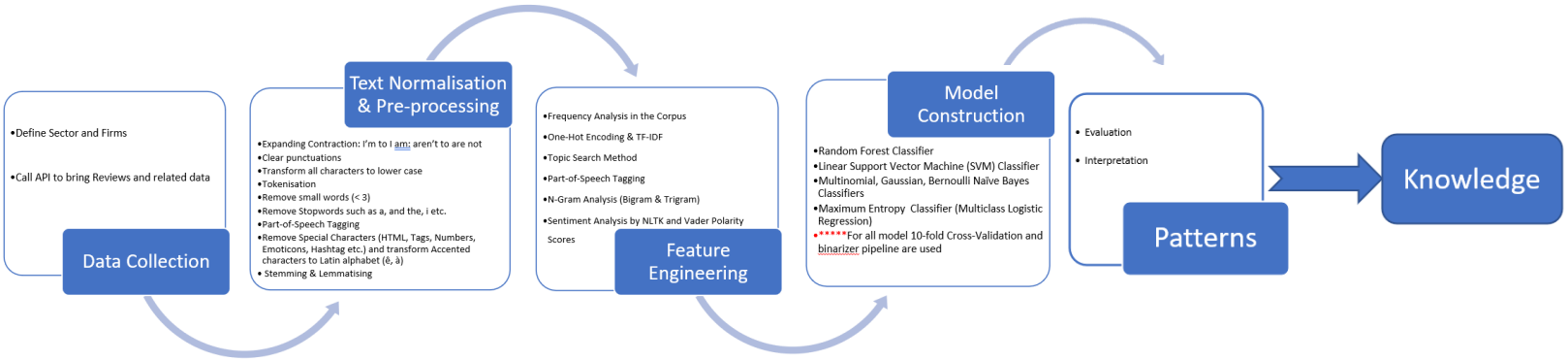

Figure 1: Brief Summary of Analysation Process

Besides all these modelling and higher-level analyses, some basic descriptive analyses on thumbsup count vs score, version average scores or review length vs score are also conducted and presented in the result part of the report.

3. RESULTS

As descriptive analyses,

- In the last 2000 reviews made by each company, extreme scores 1 and 5 are more generally used by customers to score (See Figure 2). For Amazon, 5 stars make up the vast majority, with total of 1073 reviews
- When eBay and Argos have replied to a limited number of reviews among those 2000, Amazon never return to customers for either good or bad reviews on Google Play Store (See Figure 3)
- Also, for Thumbs-up score (See Figure 5), mostly used for negative reviews, might show the number of customers who have the same issue.

Thus, not-answering behaviour can be changed with the techniques proposed in this paper, and urgent customer issues categorised by their topic might be directed to related departments to answer.

- Some analyses were conducted on versions and length of reviews (See Figure 4). When there is no significant pattern observed for versions, customers leave long texts on average for negative comments.

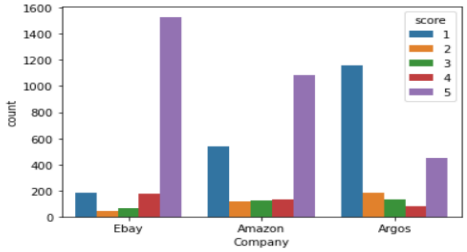

Figure 2: Analysis of Last 2000 Reviews for 3 Companies in terms of Review Scores

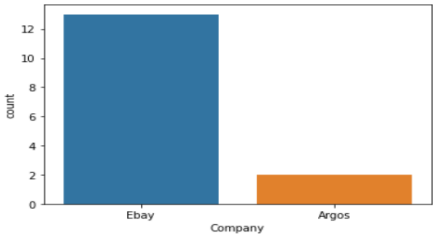

Figure 3: Replies to Last 2000 Reviews from 3 Companies

- With bags-of-world methodology, app and amazon are the top 2 most frequent words in all reviews as expected (See Appendix Figure 1)

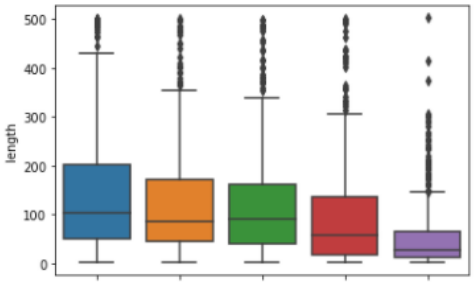

Figure 4: Length vs. Review Point Analysis

Topic Search:

- To suggest managerial insight, Negative Category Results are shared below, but analysis can be conducted for each category for the future.
- 10 topics are decided to be enough according to Figure 7 and 8
- Word clouds including top 10 popular words in each topic are created and Top 10 Bigram/Trigram that includes those words are founded (See Figure 8 below, for all topics' world cloud and topic guesses See Appendix Figure 2)
- In Figure 8 example, customers complain about the performance of our app on Samsung phones and inconvenience in Samsung mobile software. Reviews classified in Topic 4 directly are canalized to the department who are responsible for android devices and the department easily starts their investigation from Samsung phones.

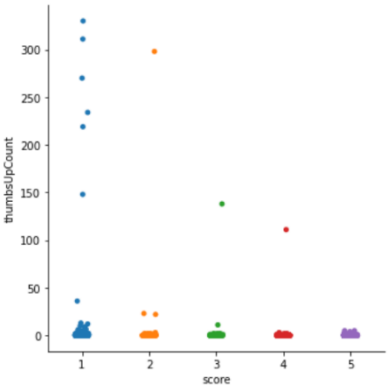

Figure 5: Thumbs-Up Scores for Each Star Group

For sentiment analysis,

- Positivity/Negativity Score of each review by NLTK is presented for each sc ore below Figures 9 and 10. Also, Vader version of this analysis can be seen in Appendix Figure 3.
- As expected, all score means are consistent with their usage, but the differen ces between consecutive scores are not separated clearly.
- Although a model to classify comments from other sources that do not includ e score criteria might be constructed on ready-to-use packages outputs, company-specific classificatio n methods can perform better to analyse and interpret.

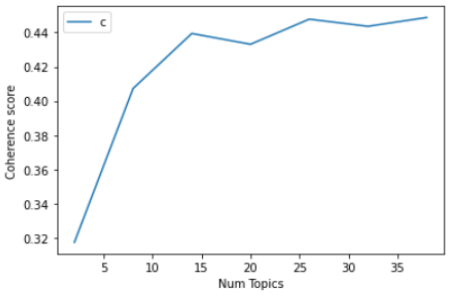

Figure 6: Coherence vs. No. Topics

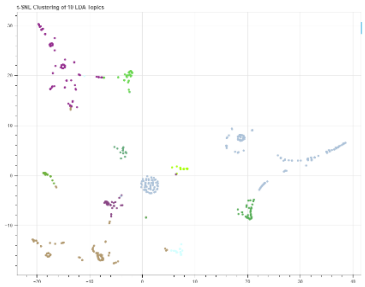
Figure 7: Cluster Representation of Topics

Bi-Gram	Tri-Gram
samsung , galaxy	account, crash, samsung
samsung , note	account, samsung, mobile
app, samsung	amazon, browser, samsung
crash, samsung	app, phone, samsung
phone, samsung	app, samsung, last
samsung, ultra	app, samsung, update
account, samsung	app, way, samsung
browser, samsung	browser, samsung, galaxy
covid, samsung	concerned, covid, samsung
day, samsung	constant, crash, samsung

Figure 8: Word Clouds contains 10 most popular word for Topic 4 and for Bigram and Trigram including 'samsung'

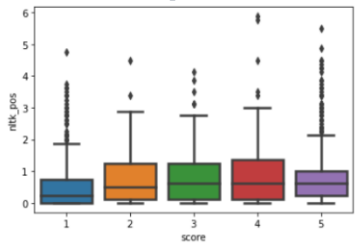

Figure 9: Positivity Score of Each Score by NLTK

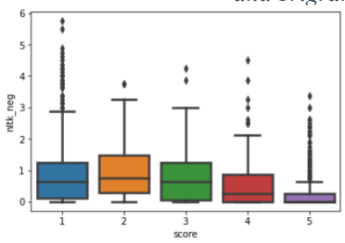

Figure 10: Negativity Score of Each Score by NLTK

For Classification Models:

- X`Best Accuracy is reached in 10-fold sets by Naïve Bayes Algorithm (See Figure 11)

- Representative example of the Naïve Bayes, the model is capable of classifying low reviews correctly by 65% (For ROC curves, Misclassified examples for each class See Appendix Figure 3 and 4).
- In the bag-of-word technique, the most frequent N-grams are presented in Figure 14. As clearly seen, people, even in the average reviews, are complaining mostly about the situation after an update.

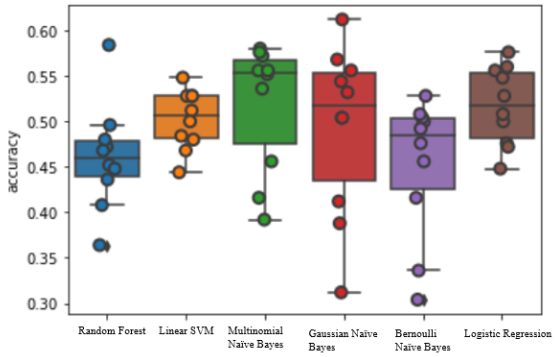

Figure 11: Model Comparison in terms of their Accuracy based on 10 crosses validated test-train samples

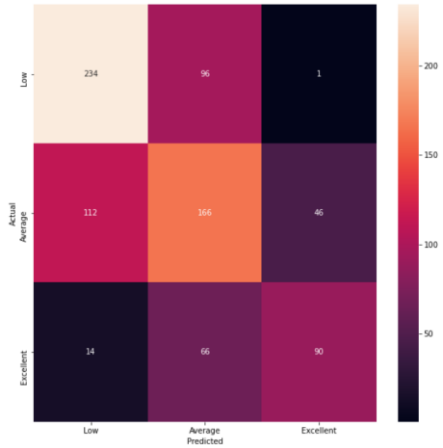

Figure 12: Sample Confusion Matrix for Naïve Bayes

	Metric			
Class	Precision	Recall	F1-Score	No. of Occurrences
Negative	0.65	0.71	0.68	331
Average	0.51	0.51	0.51	324
Positive	0.66	0.53	0.59	170

Figure 13: Classification Summary of Naïve Bayes with different metric for each category

	Most Common N-Grams		
Class	Uni-Gram	Bi-Gram	Tri-Gram
Negative	hate, horrible, account, awful, company	latest update, keeps crashing, customer service, new app, having problems	new update s***,last update app, new update horrible
Average	good, crashing, better, access, problem	good service, update crashes, app great, app open, app glitchy	good service app, new update app, last update open
Positive	love, easy, great, awasome, best	great service, great app, love amazon, excellent service, love app	app easy use, love shopping amazon, app last update

Figure 14: Most Common N-Grams of V

4. CONCLUSION AND FUTURE WORK

4.1 Conclusion

This study analyses reviews of customers in Google Play UK about the Amazon shopping app using the Natural Language Processing technique. Customers' opinions are interpreted using machine learning techniques. It is founded that even if customers give higher scores overall, customers sometimes complain about some crucial issues such as problems after an update or late delivery. The first suggestion is that after detecting the topics, the related department should immediately be alarmed to analyse the situation. For example, as in the Samsung case in the previous section, if customers talk about an inconvenience about a specific type of

brand/platform (Samsung Galaxy phone), the analyser knows where they need to start before the problem becomes more serious. Secondly, sentiment analyses should be deepened by enriching them with other sources. Lastly, as mentioned in the article of Kwakyi (2016), answering the reviews helps increase our customer satisfaction, and it also creates a positive corporate perception of the public.

4.2 Future Work

1. Tuning parameters of algorithms that are used for classification and topic search.
2. To use other sources and platforms such as tweets or Facebook comments in review classification, trained classification models will be hired (Bengfort, Bilbro, and Ojeda, 2018).

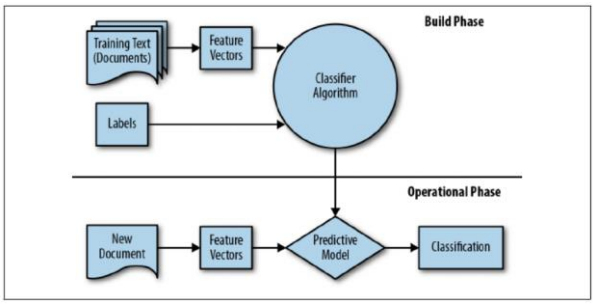

Figure 15: Text Classification Workflow (Bengfort,Bilbro, and Ojeda, 2018)

3. Polarity scores of WordNet can be used to classification by determining different thresholds (Hamouda and Rohain, 2011)
4. Online Spammers that create fake opinions should be checked and eliminated for further analysis (Fang and Zhan, 2015)
5. Working on lowercase transformation error such as US (United States) to us (pronoun) and try to minimise to meaning shifts
6. Dealing with negation for sentiment by considering a sequence of words rather than bag-of-words approaches

5. REFERENCES

Bengfort, B., Bilbro, R. and Ojeda, T. (2018). *Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning*. 1st ed. Sebastopol, CA.: O'Reilly Media Inc.

Coppola, D. (2021). "E-commerce worldwide – Statistics & Facts". *Statista, Key Figures of E-Commerce* [Online]. Available at: statista.com/topics/871/online-shopping/ (Accessed: 1 April 2021).

Fang, X., and Zhan, J. (2015). "Sentiment analysis using product review data". *Journal of Big Data*[Online]. Vol. 2(5). Available at: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2 (Accessed: 3 April 2021).

Hamouda, A., Rohaim, M. (2011). "Reviews Classification Using SentiWordNet Lexicon". *The Online Journal on Computer Science and Information Technology (OJCSIT)*. Vol. 2(1).

Kumar, K. (2018). "Evaluation of Topic Modeling: Topic Coherence". Advanced Modelling in Python, *DataScience+*[Online]. Available at: https://datascienceplus.com/evaluation-of-topic-modeling-topic-coherence/ (Accessed: 3 April 2021).

Kwakyi, G. (2016). "Why You Need to Reply to Your User Reviews". *Medium*[Online]. Available at: https://incipiagabe.medium.com/why-you-need-to-reply-to-your-apps-google-play-user-reviews-88226e617f0b (Accessed: 4 April 2021).

PYMNTS. (2020). "Exclusive Data: Twenty-Five Percent of Costumers Say Retail, Online Grocery Digital Shift Will Stick". *Coronavirus E-commerce* [Online]. Available at: https://www.pymnts.com/coronavirus/2020/the-great-reopening-doubling-down-on-digital-coronavirus-ecommerce-contactless/ (Accessed: 1 April 2021).

Redbytes. (2021). "10 Best Free Shopping Apps UK 2021"[Online]. Available at: www.redbytes.co.uk/shopping-apps-uk/ (Accessed: 2 April 2021)

Vasa, R., Hoon, L., Mouzakis, K. and Noguchi, A. (2012). "A preliminary analysis of mobile app user reviews". *Proceedings of the 24th Australian Computer-Human Interaction Conference on - OzCHI '12*', pp. 241-244[Online]. Available at: https://doi.org/10.1145/2414536.2414577 (Accessed: 2 April 2021).

5.1 Sources for Coding

Bansal, S. (2016). "Beginners Guide to Topic Modeling in Python". *Analytics Vidhya*[Online]. Available at: https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/

Li, S. (2018). "Multi-Class Text Classification with Scikit-Learn". *Towards Data Science*[Online]. Available at: https://towardsdatascience.com/multi-class-text-classification-with-scikit-learn-12f1e60e0a9f

Malinski, P. (2020). "Designing a sentiment analysis application for TripAdvisor reviews in Python". *Rk Blog- Python Programming*. Available at: https://rk.edu.pl/en/designing-sentiment-analysis-application-tripadvisor-reviews-python/

Prabhakaran, S. (2020). “Topic Modeling with Gensim (Python)”. *Machine Learning Plus*[Online]. Available at: https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/

Prabhakaran, S. (2021). “Topic modeling visualization – How to present the results of LDA models?”. *Machine Learning Plus*[Online]. Available at: https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/

Perkins, J. (2010). “TEXT CLASSIFICATION FOR SENTIMENT ANALYSIS – NAIVE BAYES CLASSIFIER”. *PYTHON NLTK COOKBOOK*. Available at: http://help.sentiment140.com/for-students/

6. APPENDIX

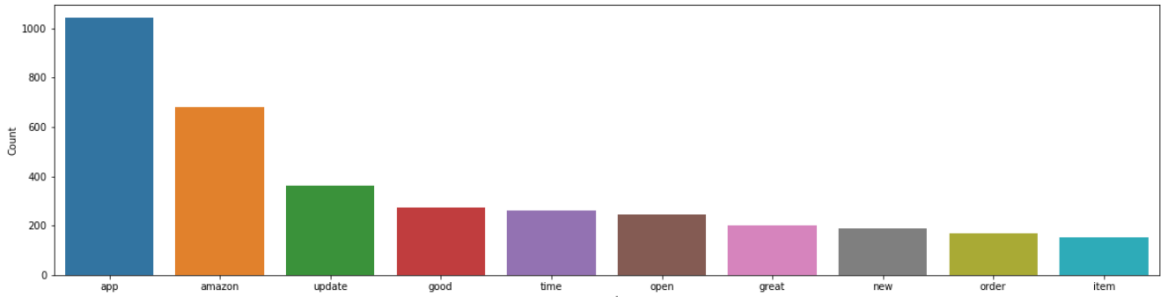

Figure 1: Basic Frequency of Words in the Word - Bag

Figure 2: Word Clouds made up of 10 most popular word for Each Topic and Topic assumptions

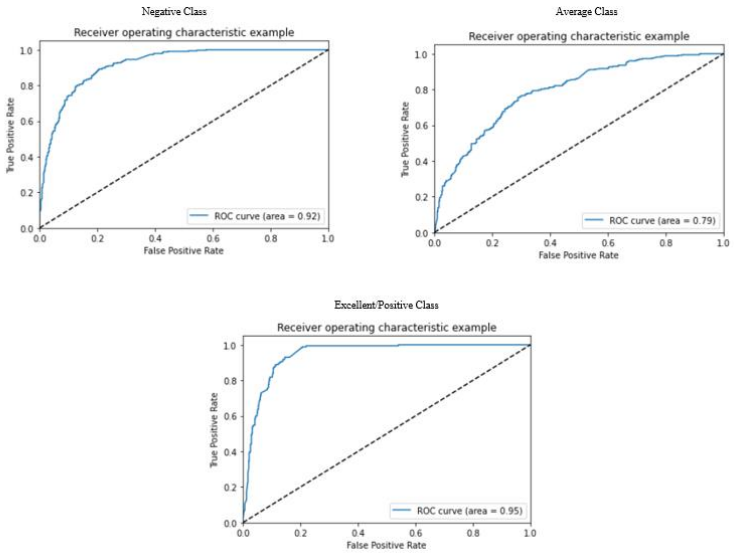

Figure 3: ROC Curves of Each Class for Naive Bayes example

	Categories				
	'Average' predicted as 'Negative'	'Excellent' predicted as 'Negative'	'Negative' predicted as 'Average'	'Excellent' predicted as 'Average'	'Average' predicted as 'Excellent'
Sentences	has its ups downs	enjoy the free trials and gives free shipping	closes soon try open	they been good always get what order time	awesome
	keep getting message that app has stopped working	cleaning tool keeps warning about this app	wont allow open app now	always convenient when needing last minute	great
	not given enough freedom use account for business	but stuff are way too expensive	always been good but since this last update	for shopping you can literally find anything	love
	app keeps crashing since this morning	the buying and return process are plain	ever since the update this application keeps	the amazon app every day whether buying	great app soooooo slow though
	addictive buying habits are spurred from this app	simple interface and functioning app	they took gift credit for idea why and not	great selection and prices	very easy order

Figure 4: Misclassified Sentences of each category

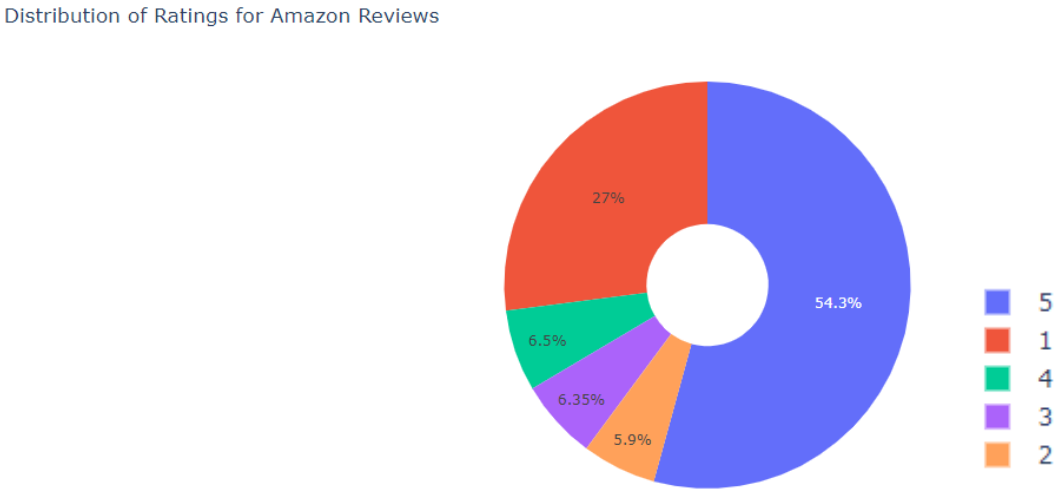

Figure 5: Distribution of Ratings for Last 2000 reviews of Amazon shopping app