Management

FORM FOR THE SUBMISSION OF ASSESSED COURSEWORK

THIS FORM MUST BE COMPLETED AND ATTACHED TO **ALL** ASSESSED WORK.

Student Number (7 digit number)	**2**	**3**	**6**	**6**	**0**	**3**	**6**		
Username (2 letters & 5 digit number ab12345)	**u**	**e**	**2**	**3**	**3**	**0**	**3**		
Unit Code	**E**	**F**	**I**	**M**	**M**	**0**	**1**	**3**	**9**
Unit Title	**Social Media and Web Analytics**								

By submitting this work online using my unique log-in and password, I hereby confirm that this work is entirely my own. I understand that all marks are provisional until ratified by the Faculty Examination Board.

Please remember to save **your work** either as a **Word document (.doc or** as a **PDF (.pdf)** unless another format is specified by the Unit Director with the filename [Unit Code_Student Number] **e.g. MGRCM0000_1234567.docx**.

Please start writing your coursework on the next page.

Sentiment Analysis of Amazon Product Reviews

Introduction

E-commerce companies like Amazon or Alibaba are a marketplace in which you may buy all the things you need in daily life. People tend to search for products through these mega-scale online shops since the categories of products are more diverse than the on-site stores. Due to the high volume of sales, the companies could collect tons of data from customers, the data may include the price of products, the quantity sold of products and the information of customers such as their reviews after shopping, genders, addresses, and income. All of the information could be translated into valuable business insights; for instance, the products with positive reviews should be placed at the top of the search list; the management must make sure the inventory of the products with the most quantity sold is sufficient. From the perspective of customers, they can enhance their understanding of a product and make informed decisions by extrapolating information from reviews (Mohammad Abu Kausar, Sallam Osman Fageeri and Arockiasamy Soosaimanickam, 2023). Thus, sentiment analysis on product reviews could be necessary for e-commerce companies.

In my study, I will use Amazon product reviews as a factor in data analysis. In the beginning, I will clean data by removing punctuations, numbers, URLs, emails, non-English words, and emojis and transforming the words to lowercase. Then, I will conduct NLP techniques like BoWs to find out which words appeared most frequently and TF-IDF to figure out which words are most important in the dataset. After this, I would draw a graph by WordCloud to see the important words in the corpus. Then, I will conduct Topic Modelling Latent Dirichlet Allocation (LDA) to look through multiple topics. Finally, I will use the Naive Bayes model and Logistic Regression model to train the dataset to make predictions and get an accuracy rate and Confusion Matrix respectively.

Literature Review

Faisal Hassan et al. (2023) investigated the effectiveness of combining various algorithms such as K-Means Clustering, Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT) on sentiment analysis. Combining each algorithm with K-Means led to notably high levels of accuracy. Specifically, the combination of K-Means with Logistic Regression (LR) achieved an accuracy rate of 99.98%. Similarly, integrating K-Means with Random Forest (RF) resulted in an accuracy of 99.906%. Lastly, merging K-Means with the Decision Tree (DT) Algorithm yielded an accuracy of 99.83%. Mahdi Rezapour (2021) examined three distinct machine learning-based algorithms, namely Naïve Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT), for sentiment classification. Performance analysis revealed that the Decision Tree algorithm surpassed the other two methods in effectiveness. The investigation from Monika, R. et al. (2019) focused on sentiment analysis utilizing the Recurrent Neural Network (RNN) model, coupled with Long-Short Term Memory Networks (LSTMs). These LSTM units are adept at handling long-term dependencies by incorporating memory within the network model, facilitating prediction and visualization. Ayyappa, Y. et al. (2023) explored the use of BERT (Bi-directional Encoder Representations from Transformers)

and LSTM (Long Short-Term Memory) for stock price forecasting. It leverages historical stock data alongside sentiment classification of Twitter posts. Lin, J., Najafabadi, M.K. (2023) presented a novel neural network model for aspect-level sentiment analysis, which integrates CNN, Bi-LSTM, and attention mechanisms. The goal is to improve the accuracy of sentiment analysis, particularly when analyzing sentences containing aspect terms. The study of Khabour, S.M. et al. (2023) performed a comprehensive emotion mining and sentiment analysis on Arabic mobile banking reviews. It utilizes machine learning, natural language processing, and resampling methods to extract subjective feedback, ascertain polarity, and discern customers' sentiments within the banking domain. Alghamdi, S. (2023) analyses customer complaints within two prominent cloud services, Microsoft Azure and DigitalOcean. The objective is to identify recurring issues faced by customers in utilizing cloud services and offer insights to cloud providers for enhancing customer satisfaction. The study by Agarwal, A. (2020) concentrated on sentiment classification and examined its impact on changes in stock market prices. It provides investment insights by employing sentiment analysis, utilizing the VADER (Valence Aware Dictionary and Sentiment Reasoner) tool, on a selection of the most liquid stocks. Nagaraj, P. et al. (2023) illustrated the efficacy of sentiment analysis techniques in predicting the sentiment of movie reviews. These findings hold potential applications across diverse fields, including movie recommendation systems, market research, and opinion mining.

The research of sentiment analysis is plenty and robust. The research above shows that there are many machine-learning and deep-learning algorithms, K-Means, Decision Trees, Random Forest, SVM, RNN, LSTM, CNN, BERT, and specific combinations of algorithms, applied in this field. Besides, sentiment analysis can be used in multiple areas; for instance, we can apply sentiment analysis to movie reviews, predict equity prices, evaluate banking services, identify consumer complaints, and analyse financial news. In my study, I will conduct topic modelling LDA to do sentiment analysis, then I will interpret my findings based on the results, and I will use Logistic Regression and Naive Bayes Classifier models to do predictions in the end.

Methodology and Data

Data Collection and Cleaning

In my study, I used the dataset from Kaggle (See Appendix), the dataset contains Amazon Product Reviews, in which there are two columns: ‘Text’ and ‘Label’, the shape of the dataset is (19,996, 2). I will not use the ‘Label’ column, since there are only two values: 1 for ‘positive’ and 0 for ‘negative’, and 15,230 and 4,766 values respectively. I would like to use another method to dig deeper into the reviews, so I will divide the reviews into three values: ‘positive’, ‘negative’ and ‘neutral’.

First, I pre-processed the texts by removing Emails, URLs, punctuation, emojis, and non-English characters and transforming texts to lowercase. Then, I saved these pre-processed texts to column ‘Review’. After removing the reviews which are less than one word and removing duplicates, I got 19,983 values remaining. Second, I lemmatised the reviews and removed the default stopwords.

Sentiment Analysis

I applied Bag of Words (BoW) analysis and TF-IDF analysis and plotted the figures. Figure 1 shows the result of BoW analysis, and we can find out the top 10 most frequent words are ‘app’ (10,664 times), ‘game’ (6,722 times), ‘love’ (4,264 times), ‘great’ (4,018 times), ‘like’ (4,015 times), ‘get’ (3,466 times), ‘wa’ (3,417 times), ‘use’ (3,251 times), ‘kindle’ (2,889 times), time (2,877 times). The word ‘wa’ refers to the word ‘was’, the word is lemmatised. From the result, we can see that the reviews of Amazon products may focus on the consumer experience after shopping. We can interpret that most reviews are about how great is Kindle, people spend much time on it and really like/ love it, and maybe there are many game apps on the Kindle. Figure 2 shows the result of the TF-IDF analysis, and we can see that the top 10 words with the highest TF-IDF scores are helpful, like, fun, love, favourite, small, cool, hi, moo, bang, and all of them score 1. According to the outcome, most of the words are positive such as ‘helpful’, ‘like’, ‘fun’, ‘love’, ‘favorite’, and ‘cool’, maybe we can interpret that people love to use Kindle and think it is helpful; besides, the word ‘moo’ may be sound of a game, people may think the games on Kindle are fun and cool.

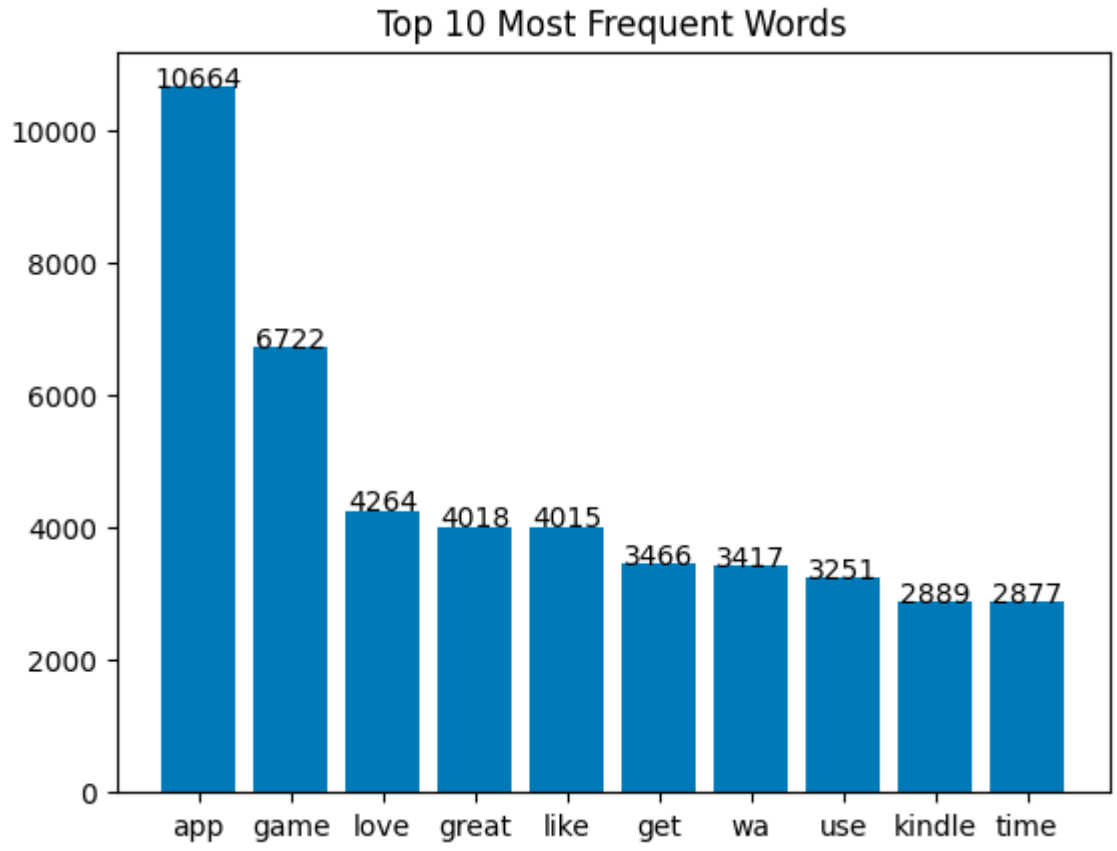

Figure 1: Top 10 Most Frequent Works by BoW

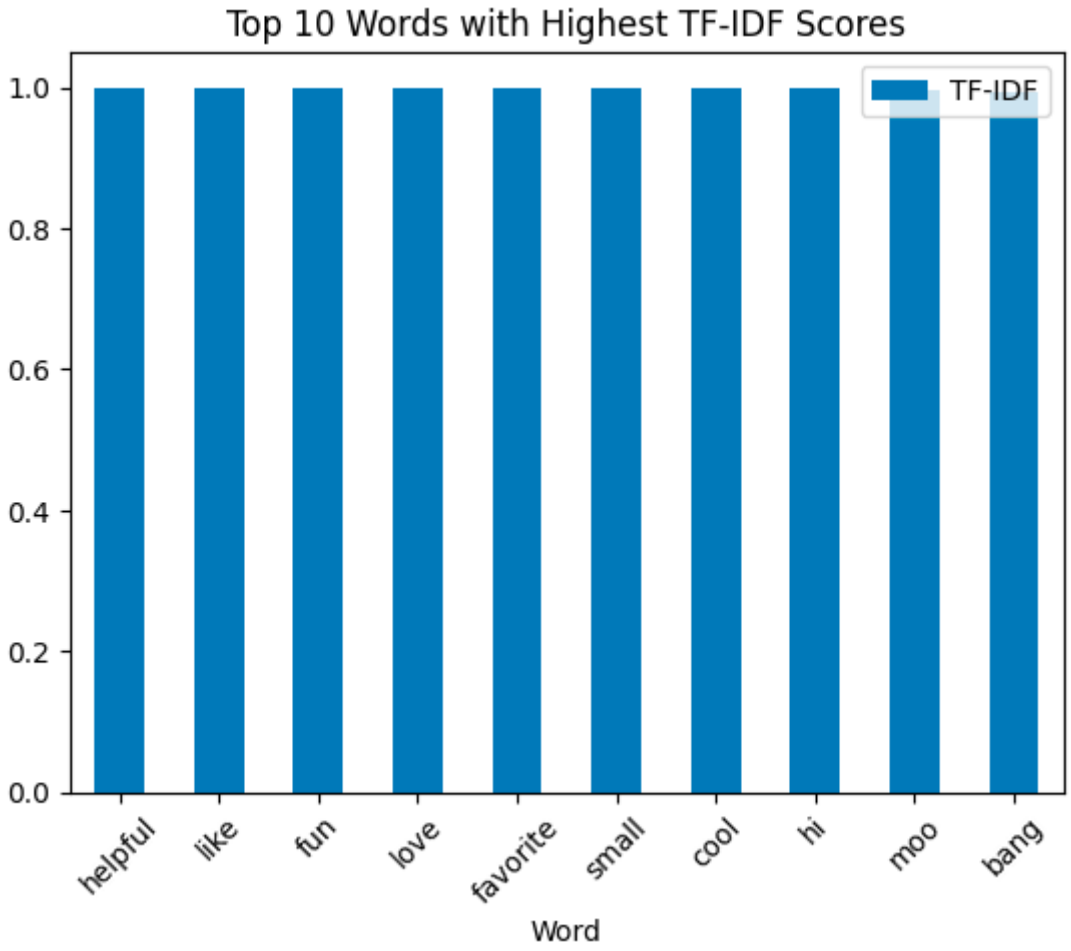

Figure 2: Top 10 Words with Highest TF-IDF Scores

After getting the single-word results, I applied TF-IDF and Word2Vec models to recommend ten sentences from the one I chose. I calculated the Euclidean distance of review vectors to see the similarity between sentences. From the results of Figures 3 & 4, we can see an interesting result: the top 2 recommended sentences from the TF-IDF model are the bottom 2 recommended sentences from the Word2Vec model. Besides, the Euclidean similarity scores of Word2Vec are higher than those of TF-IDF. We may interpret that the Word2Vec model is better at finding similarity, it can find precise similar terms in the document. Besides, we can tell that the Word2Vec model is better by comparing the chosen sentences from each model. The queried review is “really like game easy play close board game get hooked”, the top 1 recommended sentence from TF-IDF is “really fun game bored get game u play…”; and the top 1 from Word2Vec is “really enjoy playing game make use brain like…”. The queried review is talking about how he or she likes the game, the sentence from TF-IDF contains the word “bored”, which is a opposite meaning from the sentence. So, the sentence selected from Word2Vec is closer to the queried review.

```
============================== Queried Review Details ==============================
Review: really like game easy play close board game get hooked

 ======================== Recommended Reviews: ========================
```

	Review	Euclidean similarity
1	really liked concept using mirror prism reflec...	0.554790
2	amazing game play totally addicting play story...	0.565091
3	grandkids play cant get anywhere game require ...	0.575762
4	game funniest game ever heard really get broth...	0.587819
5	easy play understand character cute game color...	0.603348
6	fantastic game challenging different game play...	0.604801
7	really enjoyed playing game easy learn play se...	0.605046
8	simple memory game picture different yet simil...	0.608624
9	game similar old atari pong excellent graphic ...	0.624839
10	really fun game bored get game u play long tim...	0.627972

Figure 3: Recommended Reviews from the TF-IDF model

```
============================== Queried Review Details ==============================
review: really like game easy play close board game get hooked

 ======================== Recommended Reviews: ========================
```

	Review	Euclidean similarity
1	really fun game bored get game u play long tim...	0.569727
2	game similar old atari pong excellent graphic ...	0.635790
3	like stop playing time pick right leave good g...	0.651026
4	game breakout offer many twist change make ori...	0.654627
5	play game lot great fun play computer game cha...	0.656154
6	play game every day would recommend everyone l...	0.656380
7	game great age sad know older people hooked ga...	0.658190
8	game fun play really get hooked find wanting p...	0.659070
9	game really fun like different style play dont...	0.662026
10	really enjoy playing game make use brain like ...	0.666768

Figure 4: Recommended Review from Word2Vec model

For constructing the sentiment, I used SentimentIntensityAnalyzer to generate scores for each review, then I classified ‘positive’ as a score equal to or higher than 0.05, ‘negative’ as a score equal to or lower than -0.05, ‘neutral’ as the remaining. From Figure 5, we can see that most of the reviews scored more than 0, which means that most reviews are positive. From Figure 6, we know that there are 15,438 positive reviews, 3,673 negative reviews and 872 neutral reviews. Besides, the original number of positive reviews is 15,230, which means SentimentIntensityAnalyzer is a feasible tool for classifying sentiment.

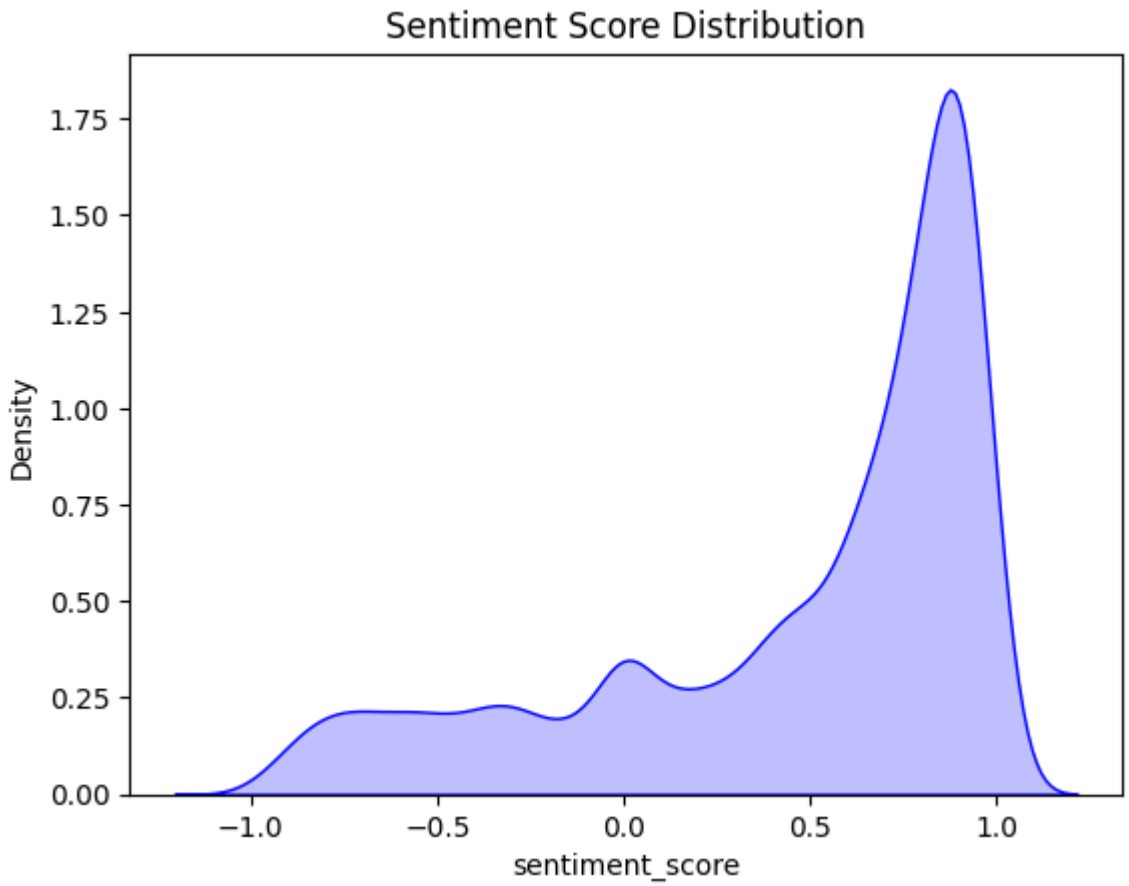

Figure 5: Sentiment Score Distribution

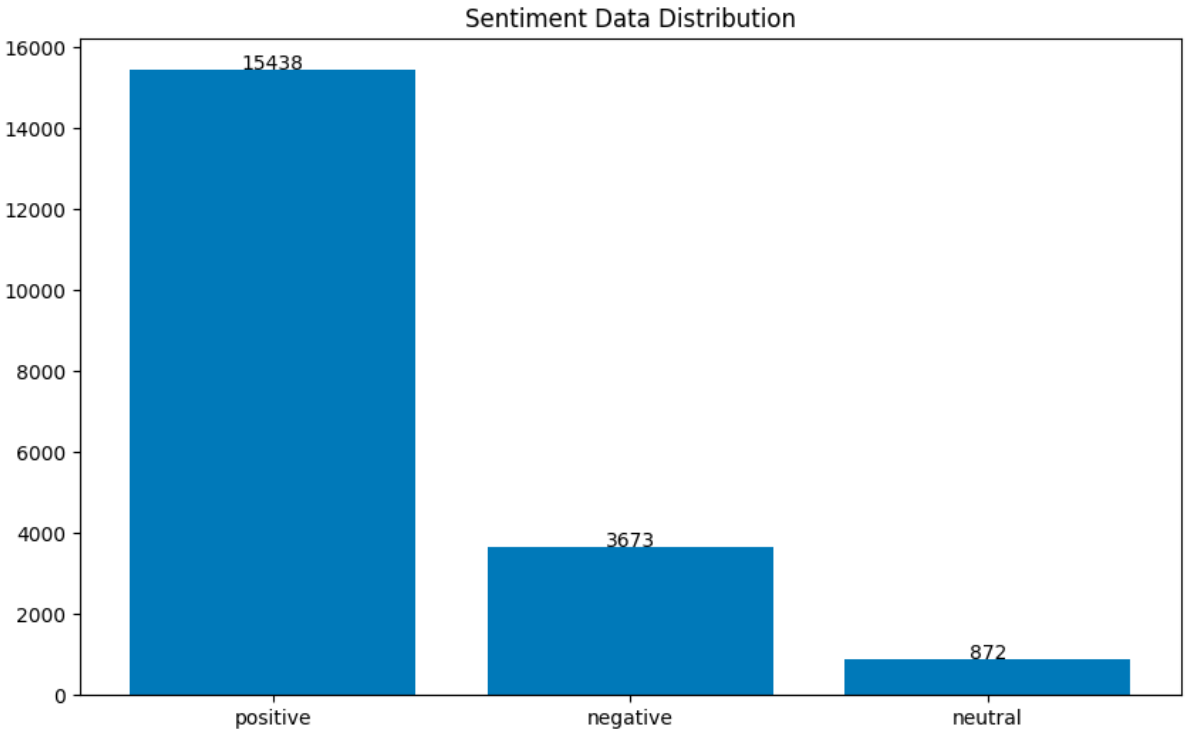

Figure 6: Sentiment Distribution (‘positive, ‘negative’ and ‘neutral’)

I applied WordCloud to visualise the frequency of words from positive, negative and neutral reviews. After getting the word clouds (See Appendix Figure7 ~9), I found out that some words like ‘wa’, ‘even’, ‘u’, ‘got’, ‘day’, ‘ha’ and ‘doe’ do not have much meaning. Thus, I used StopWords to eliminate these words and plotted revised wordclouds. From Figure 10, we can see that the most frequent positive words are ‘app’, ‘great’, ‘game’, ‘love’ and ‘kindle’, which may mean that people have positive feelings about using Kindle for playing games. From Figure 11, we can know that the negative keywords are ‘app’, ‘game’, ‘fire’, ‘dont’ and ‘kindle’; there are three words ‘app’, ‘game’, and ‘kindle’ overlapped with positive keywords, which means people have the similar reason for negative feelings, people do not like the games on Kindle, and we can know the reasons may be download problems, especially on phone, since there are negative keywords ‘download’, ‘problem’ and ‘phone’. From Figure 12, the neutral keywords are ‘time’, ‘use’, ‘app’, ‘game’, and ‘dont’, which may mean that the download time on the game app is too long, so people do not have positive feedback; besides, there is another keyword ‘android’, so maybe the phones with Android system are likely to have a download problem.

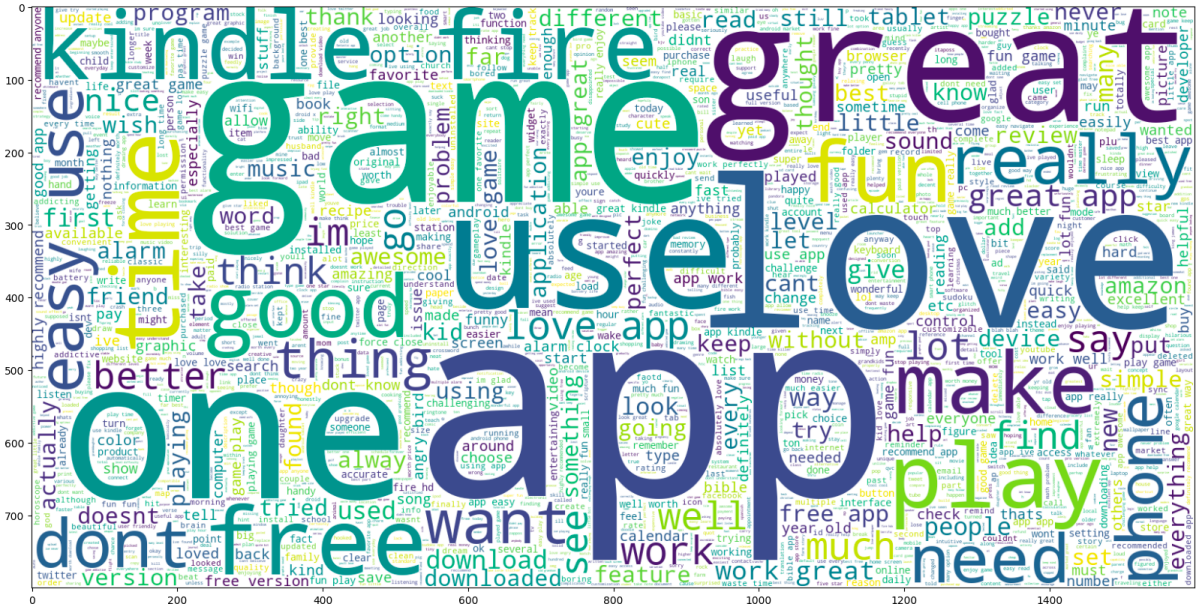

Figure 10: The word cloud of positive reviews

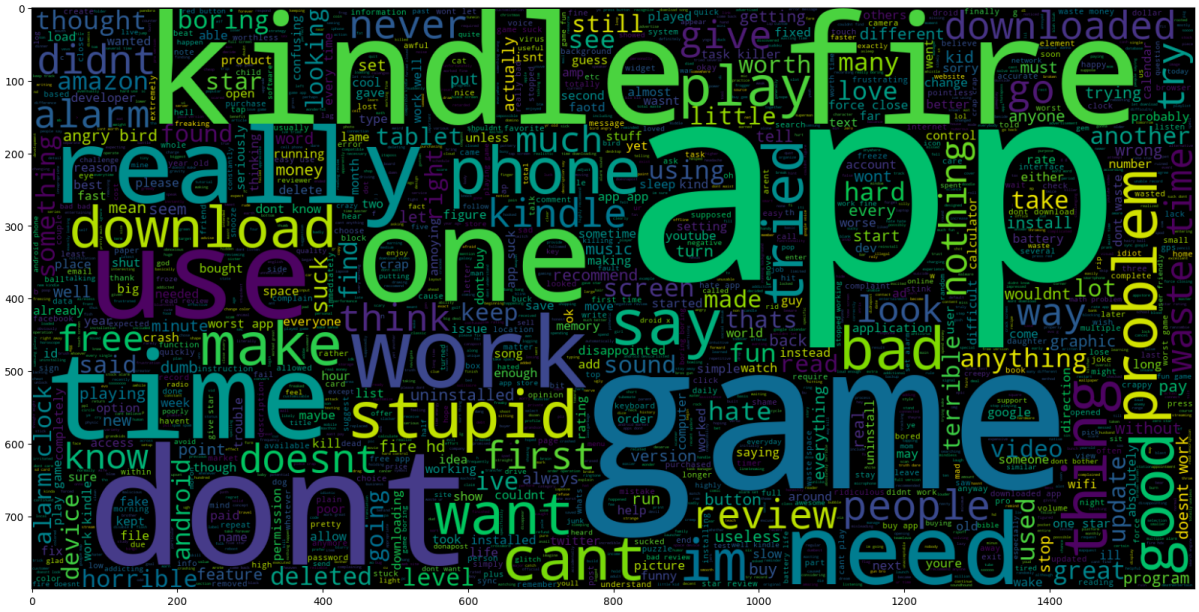

Figure 11: The word cloud of negative reviews

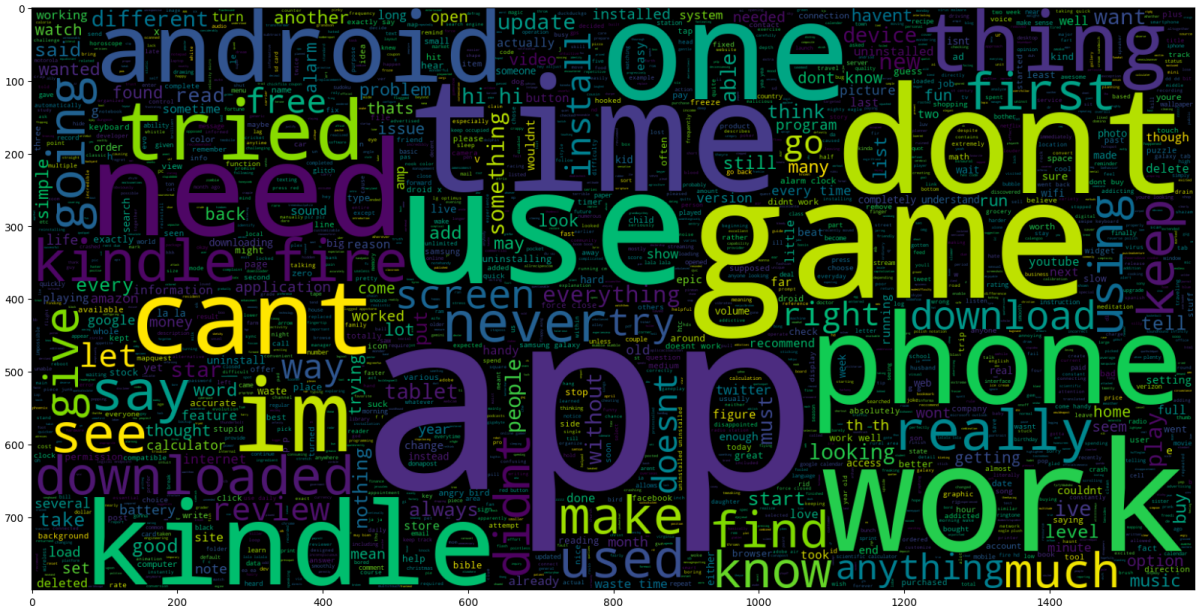

Figure 12: The word cloud of neutral reviews

LDA model

Latent Dirichlet Allocation (LDA), a widely adopted unsupervised generative probabilistic approach for modelling a corpus, is the predominant method in topic modelling. LDA operates under the assumption that every document can be depicted as a probabilistic distribution across latent topics. Furthermore, it assumes that the topic distribution across all documents follows a common Dirichlet prior. Each latent topic within the LDA model is also depicted as a probabilistic distribution across words, with the word distributions of topics sharing a common

Dirichlet prior (Jelodar, H. et al., 2019). Applying LDA to sentiment analysis of user reviews can enhance e-commerce platforms by providing insights into product quality. By analysing user reviews, sentiment analysis aims to assist e-commerce in understanding user sentiments, thereby informing potential buyers about product quality through ratings provided by users (Wahyudi, E. and Kusumaningrum, R., 2019).

Before applying the LDA model, I need to tokenise the reviews. I chose 10 topics and 30 terms in each topic to see the results and selected Metric Multidimensional Scaling ('mmds') as the dimension reduction method. From topic 1, we can see that 'app', 'game', 'kindle', 'free', and 'great' appear in the topic, which implies that people love to play the game on Kindle, and it is free of charge. From topic 2, the keywords are 'use', 'alarm', 'easy', 'time' and 'clock', which may mean that people can set a certain amount of time for playing the game, and the operation of doing it is easy. From topic 3, the keywords are 'game', 'fun', 'kid', and 'graphic', which may mean the game is fun, graphic, and especially for kids. Table 1 shows the keywords from each topic and interpretation from the keywords.

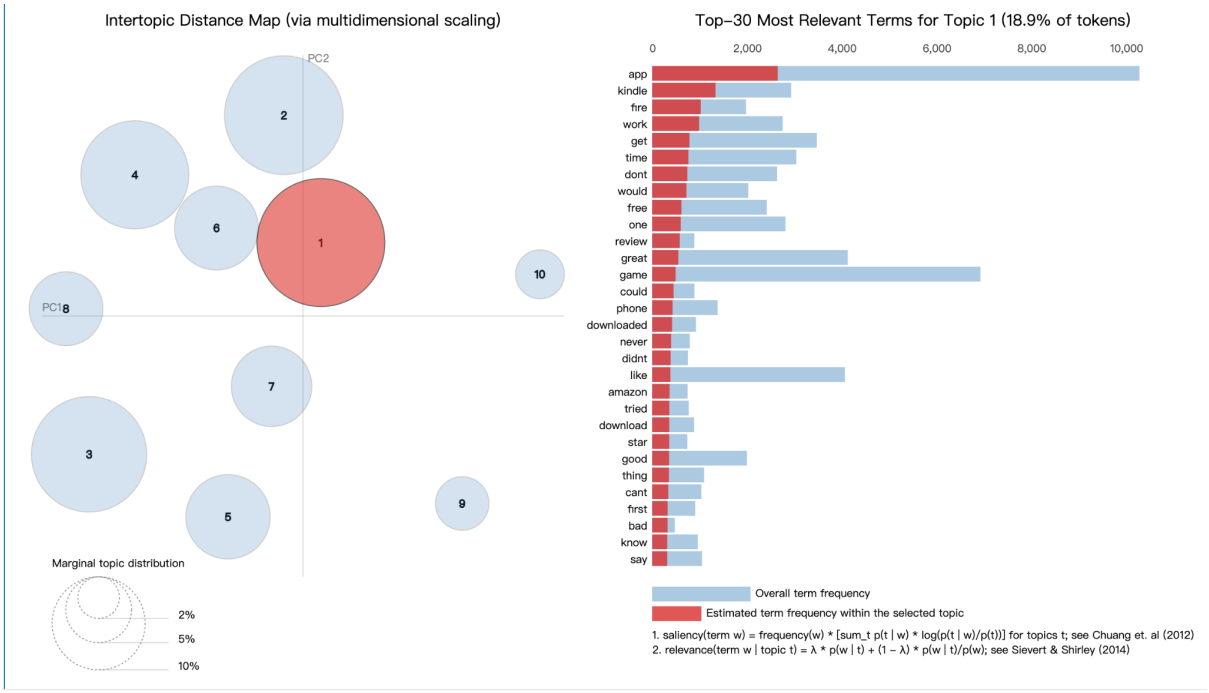

Figure 13: LDA model for topic 1

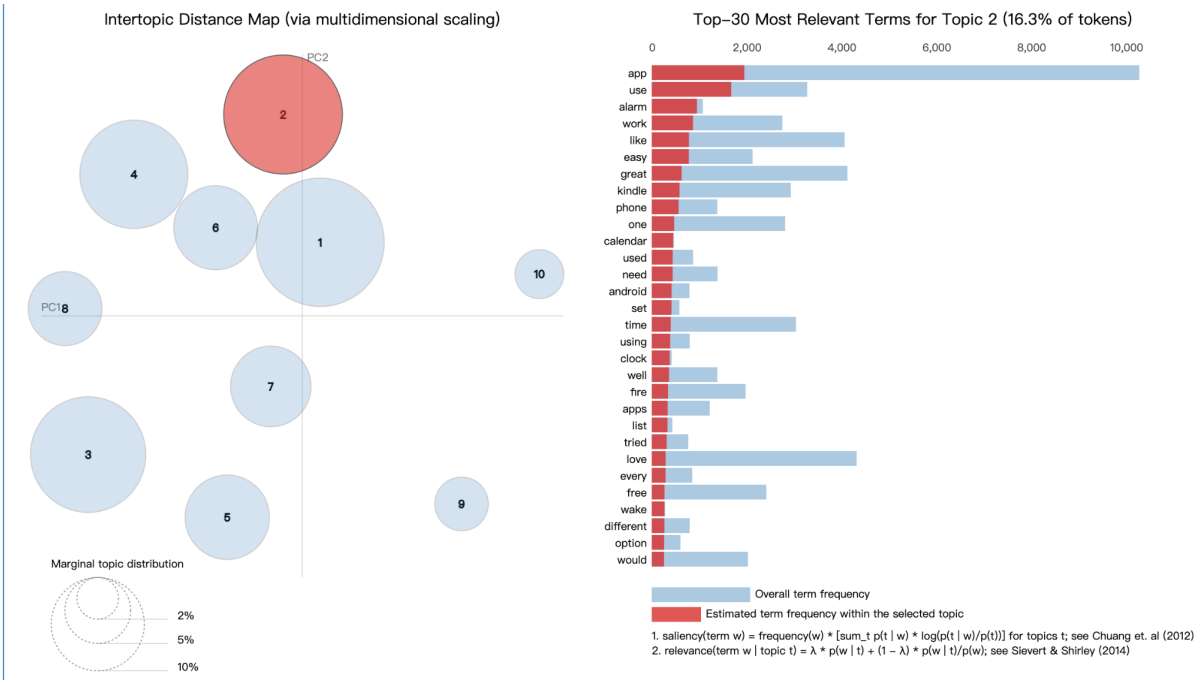

Figure 14: LDA model for topic 2

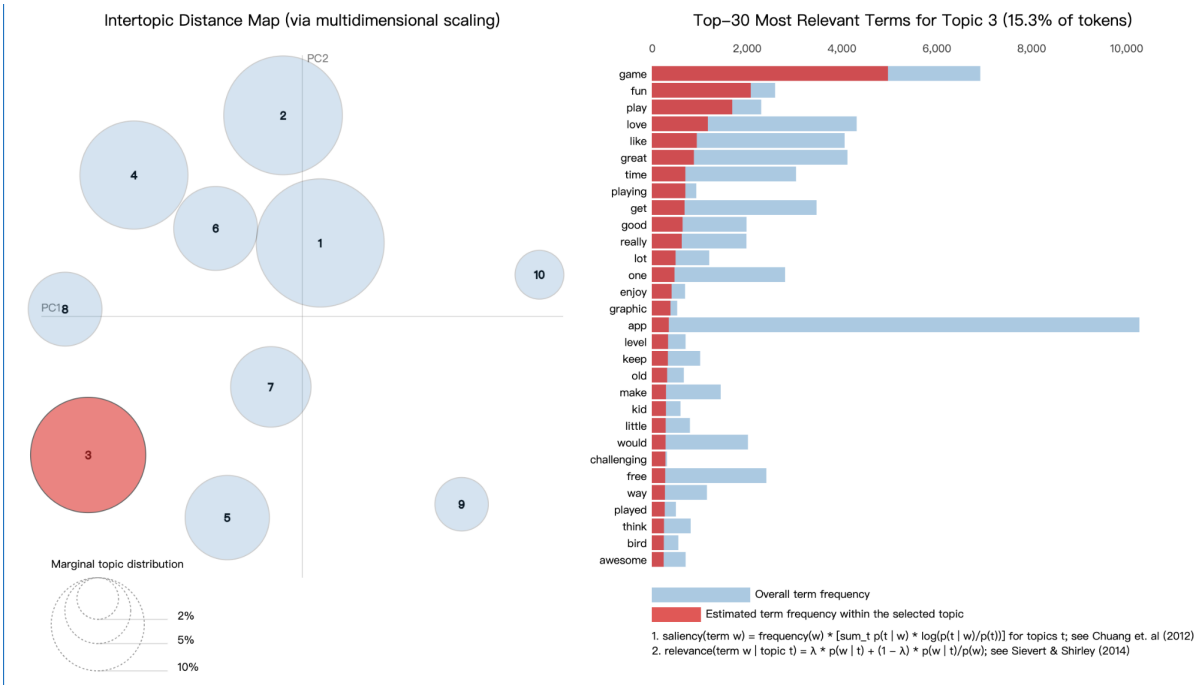

Figure 15: LDA model for topic 3

Table1:

	Keywords	**Interpretation**
Topic 1	‘app’, ‘game’, ‘kindle’, ‘free’, ‘great’	The free game on Kindle is great.
Topic 2	‘use’, ‘alarm’, ‘easy’, ‘time’, ‘clock’	An alarm in the game can make time management easy.
Topic 3	‘game’, ‘fun’, ‘kid’, ‘graphic’	The graphic and fun game is for kids.
Topic 4	‘version’, ‘recommend’, ‘highly’, ‘game’	The version of the game is highly recommended.
Topic 5	‘dont’, ‘waste’, ‘hate’, ‘boring’	The game is boring and will waste people’s time.
Topic 6	‘phone’, ‘screen’, ‘like’	People like to play the game on the phone screen.
Topic 7	‘get’, ‘tweet’, ‘know’	People know the game from tweets.
Topic 8	‘game’, ‘need’, ‘really’, ‘people’	People need the game.
Topic 9	‘video’, ‘song’, ‘buy’, ‘download’, ‘money’, ‘youtube’	People need to buy the song of the game from YouTube.
Topic 10	‘calculator’, ‘news’, ‘map’, ‘app’	The other applications on Kindle.

Predictions and Results

I would conduct two algorithms, Naive Bayes Classifier and Logistic Regression, to predict the sentiment. Then, I would present the confusion matrix and accuracy rate, precision rate, recall rate and f1 score from these two methods.

Native Bayes Classifier:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|X) = P(x_1|c) \times P(X_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

- $P(c|x)$ is the posterior probability of class (target) given predictor (attribute).
- $P(c)$ is the prior probability of class.
- $P(x|c)$ is the likelihood which is the probability of a predictor given class.
- $P(x)$ is the prior probability of the predictor.

Logistic Regression:

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}}$$

- $P(y = 1|X)$ is the probability that the outcome y is equal to 1 given the input features X.
- β_0 is the intercept term.
- β_1,β_2, …, β_n are the coefficients corresponding to the input features X_1, X_2, …, X_n respectively.
- X_1, X_2, …, X_n are the input features.
- e is the base of the natural logarithm (Euler's number).

First, I defined 'X' as reviews and 'y' as sentiment and divided 80% data for training and 20% data for testing. Second, I made matrices containing the token counts with 2-gram tokens and a maximum of 500,000 features. Third, I presented Table 2 containing compares accuracy rates from models and confusion matrices of each model. From the result of Table 2, we can tell that Naive Bayes Classifier is better on training accuracy, and the other rates are the same or very close. The result means these two models have no difference in predicting the review sentiment, and the testing accuracy is over 80%, which means these two models are both good methods for tackling this predicting problem. From Figure 16 to 17, I presented the confusion matrix, number 0 stands for 'positive', number 1 is 'negative' and number 2 is 'neutral'. As we can see, in Naive Bayes Classifier, the accuracy rate of 'positive' is 81.5% (3,019 /3,706) of 'negative' is 64.3% (184 /286) of 'neutral' is 0%; in Logistic Regression model, the accuracy rate of 'positive' is 81.0% (3,029 /3,739) of 'negative' is 66.5% (171/ 257) of 'neutral' is 0%. These results show that the accuracy rate decreases dramatically when the amount of data is few. The models cannot predict well on 'neutral' sentiment because the data of the sentiment is short. The amount of 'neutral' data accounts for 4.4% of the whole dataset (Figure 6). The models work well only in positive reviews, maybe it is because the dataset is imbalanced, thus models would like to predict the positive reviews.

Table 2: Multiple Indicators from Naive Bayes Classifier & Logistic Regression Model

	Training Accuracy	Testing Accuracy	Precision Rate	Recall Rate	F1 Score
Naive Bayes Classifier	0.988	0.801	0.747	0.801	0.754
Logistic Regression	0.999	0.801	0.747	0.801	0.750

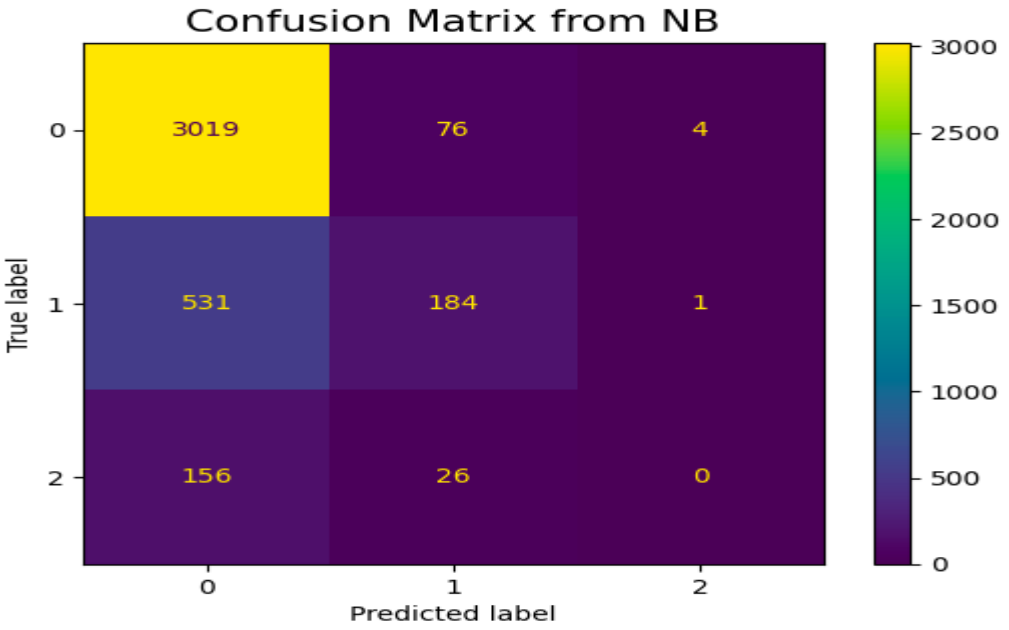

Figure 16: Confusion Matrix from Naive Bayes Classifier

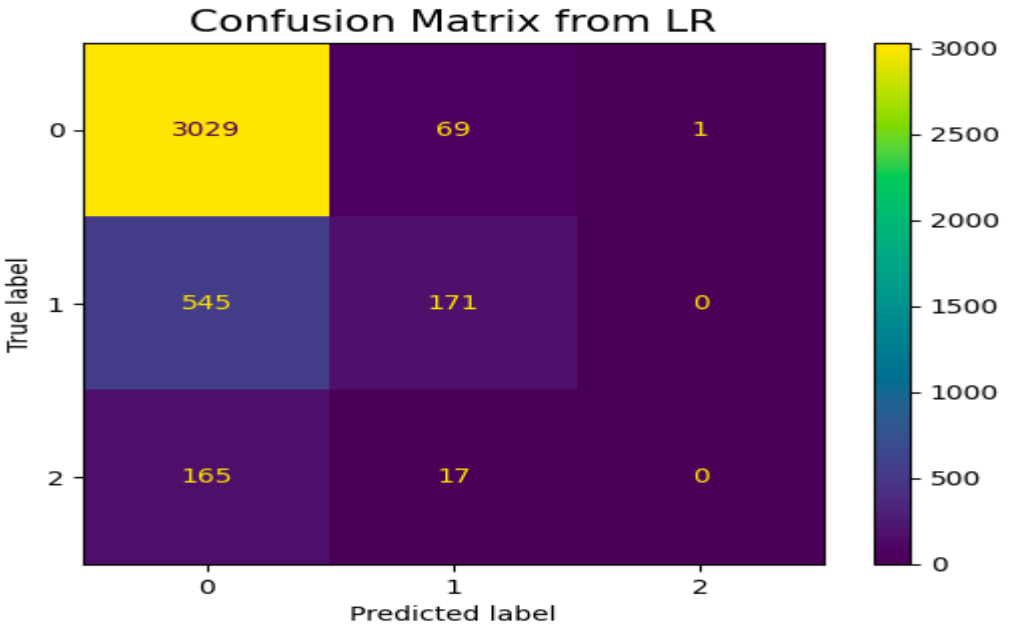

Figure 17: Confusion Matrix from Logistic Regression

Conclusion and Limitations

In this study, I analysed the Amazon Product Reviews dataset, I found that the dataset is very imbalanced, and most reviews are positive, so I used BoW and TF-IDF methods to see the frequent words and keywords. In this analysis, I had a result that the data may be reviews of a game on Kindle, most reviews are related to positive experiences, and the negative reviews are about the download problems. Then, I applied WordCloud to visualise the keywords form 'positive', 'negative' and 'neutral', I found the keyword 'android' on neutral reviews, maybe the download problem happened on Android phones. From the LDA analysis, I concluded 10 topics with 30 terms of each. From the keywords of topics, I found out that the game is for kids mainly and maybe there are other applications like calculators on Kindle. In the end, I chose Naive Bayes Classifier and Logistic Regression to train the dataset and make predictions. The result is that the overall accuracy is 80.1%, which seems good enough, but I found out that the model is good for predicting positive reviews but bad for predicting negative and neutral reviews. To sum up, the game on Kindle is great and successful for most customers, and the game is designed for kids. However, my model needs to be adjusted, maybe use another algorithm or try to make the dataset more balanced.

There are some limitations in my study. First, the dataset is very imbalanced, the positive reviews are way more than other reviews, which would cause some problems in the training model. The accuracy rate would be poor when the model predicts non-positive reviews. It is better to select a balanced dataset, and the accuracy rate may increase by doing so. Second, the dataset may only focus on one game, so the results would be limited and would make the model perform badly on similar review datasets; for instance, when there are new positive words like 'comfortable', and 'healthy' as inputs, model may predict these words as neutral or negative words. Third, most user experience may be based on Kindle or Kindle apps, so the outcome would be limited as well. If people play the game on different devices or apps, maybe they will have a different experience. Thus, my research model may only be feasible for the game on Kindle; it's challenging to adapt it for use in multiple fields.

References

Mohammad Abu Kausar, Sallam Osman Fageeri and Arockiasamy Soosaimanickam (2023) "Sentiment Classification based on Machine Learning Approaches in Amazon Product Reviews," Engineering, Technology & Applied Science Research, 13(3). Available at: https://doi.org/10.48084/etasr.5854.

Faisal Hassan et al. (2023) "Performance evolution for sentiment classification using machine learning algorithm," Journal of Applied Research in Technology & Engineering, 4(2), pp. 97–110. Available at: https://doi.org/10.4995/jarte.2023.19306.

Mahdi Rezapour (2021) "Sentiment classification of skewed shoppers' reviews using machine learning techniques, examining the textual features," Engineering Reports, 3(1), p. n/a–n/a. Available at: https://doi.org/10.1002/eng2.12280.

Monika, R. et al. (2019) "Sentiment Analysis of US Airlines Tweets Using LSTM/RNN," in 2019 IEEE 9th International Conference on Advanced Computing (IACC), pp. 92–95. Available at: https://doi.org/10.1109/IACC48062.2019.8971592.

Ayyappa, Y. et al. (2023) "Forecasting Equity Prices using LSTM and BERT with Sentiment Analysis," in 2023 International Conference on Inventive Computation Technologies (ICICT), pp. 643–648. Available at: https://doi.org/10.1109/ICICT57646.2023.10134443.

Lin, J., Najafabadi, M.K. and 2023 IEEE/ACIS 8th International Conference on Big Data, Cloud Computing, and Data Science (BCD) Hochimin City, Vietnam 2023 Dec. 14 - 2023 Dec. 16 (2023) "Aspect Level Sentiment Analysis with CNN Bi-LSTM and Attention Mechanism," in 2023 IEEE/ACIS 8th International Conference on Big Data, Cloud Computing, and Data Science (BCD), pp. 282–288. Available at: https://doi.org/10.1109/BCD57833.2023.10466355.

Khabour, S.M. et al. (2023) "Arabic Sentiment Analysis of Mobile Banking Services Reviews," in 2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS), pp. 1–8. Available at: https://doi.org/10.1109/SNAMS60348.2023.10375427.

Alghamdi, S. and 2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS) Abu Dhabi, United Arab Emirates 2023 Nov. 21 - 2023 Nov. 24 (2023) "Toward Identifying Customer Complaints of Cloud Service Providers Using Topic Modeling and Sentiment Analysis," in 2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS), pp. 1–6. Available at: https://doi.org/10.1109/SNAMS60348.2023.10375466.

Agarwal, A. and 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN) Bhimtal, India 2020 Sept. 25 - 2020 Sept. 26 (2020) "Sentiment Analysis of Financial News," in 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN), pp. 312–315. Available at: https://doi.org/10.1109/CICN49253.2020.9242579.

Nagaraj, P. et al. (2023) “Movie Reviews Using Sentiment Analysis,” in 2023 International Conference on Computer Communication and Informatics (ICCCI), pp. 1–6. Available at: https://doi.org/10.1109/ICCCI56745.2023.10128450.

Jelodar, H. et al. (2019) “Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey,” Multimedia Tools and Applications : An International Journal, 78(11), pp. 15169–15211. Available at: https://doi.org/10.1007/s11042-018-6894-4.

Wahyudi, E. and Kusumaningrum, R., 2019, October. Aspect based sentiment analysis in E-commerce user reviews using Latent Dirichlet Allocation (LDA) and Sentiment Lexicon. In *2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)* (pp. 1-6). IEEE.

Appendix

Dataset: https://www.kaggle.com/datasets/mahmudulhaqueshawon/amazon-product-reviews/data

GitHub: https://github.com/youweicheng/Web_Analytics

Figure 7: The word cloud of positive reviews (Before removing new stopwords)

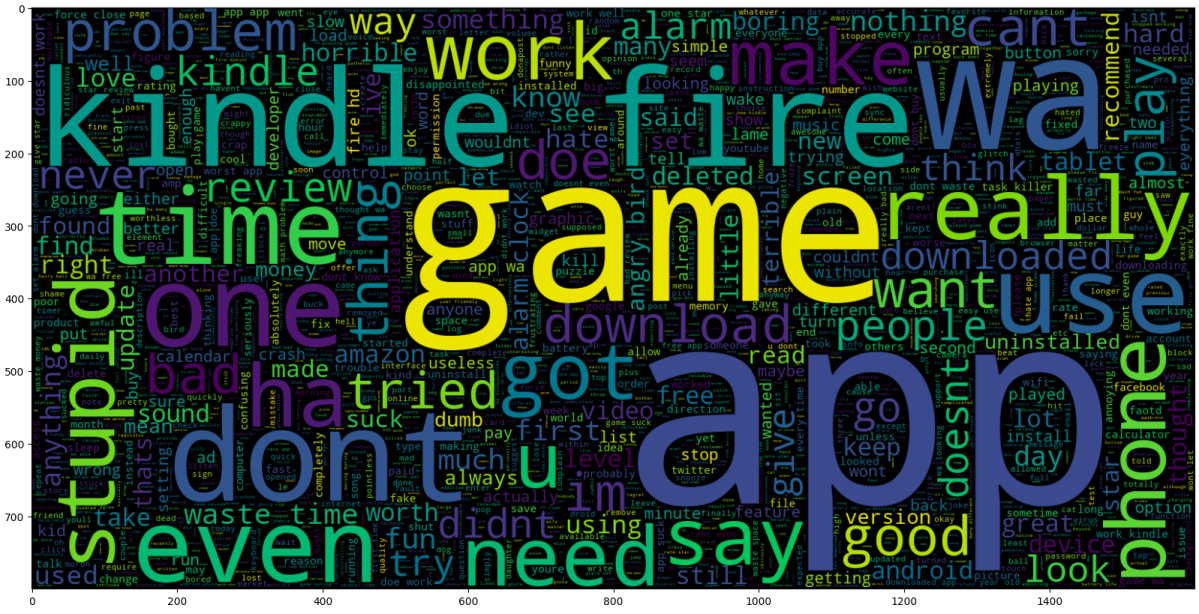

Figure 8: The word cloud of negative reviews (Before removing new stopwords)

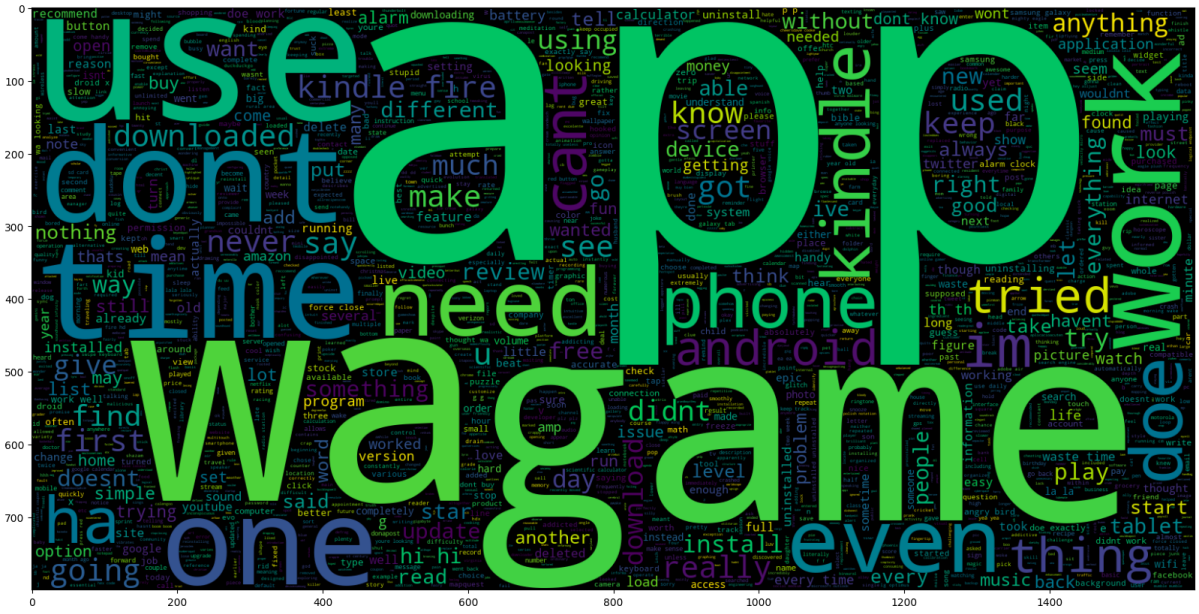

Figure 9: The word cloud of neutral reviews (Before removing new stopwords)