APPLICATION OF SENTIMENT ANALYSIS IN E-COMMERCE RECOMMENDATION SYSTEMS

ONG WEI AUN

A thesis submitted in fulfilment of the requirements
for the award of the degree of
MASTER OF SCIENCE IN DATA SCIENCE AND BUSINESS ANALYTICS

ASIA PACIFIC UNIVERSITY OF TECHNOLOGY & INNOVATION (APU)

APRIL 2022

DECLARATION OF THESIS CONFIDENTIALITY

Author's full name:	**ONG WEI AUN**
IC No./Passport No.:	**961015-07-5295**
Thesis/Project title:	**APPLICATION OF SENTIMENT ANALYSIS IN E-COMMERCE RECOMMENDATION SYSTEM**

I declare that this thesis is classified as:

☐ CONFIDENTIAL

☐ RESTRICTED

☑ OPEN ACCESS

I acknowledged that Asia Pacific University of Technology & Innovation (APU) reserves the right as follows:

1. The thesis is the property of Asia Pacific University of Technology & Innovation (APU).
2. The Library of Asia Pacific University of Technology & Innovation (APU) has the right to make copies for the research purpose only.
3. The Library has the right to make copies of the thesis for academic exchange.

Author's Signature: 

Date: 24 April 2022

Supervisor's Name: **MR. RAHEEM MAFAS**

Date: 24 April 2022

Signature: 

DECLARATION OF SUPERVISOR(S)

"We hereby declare that We have read this thesis and in our opinion, this thesis is sufficient in terms of scope and quality for the award of the degree of

Master of Science in Data Science and Business Analytics"

Name of Supervisor:	**MR. RAHEEM MAFAS**
Signature:	RMF ……
Date:	24 April 2022
Name of Supervisor (II)	**PROF. DR. R. LOGESWARAN**
Signature:	……LGR……
Date:	24 April 2022

DECLARATION OF ORIGINALITY AND EXCLUSIVENESS

"I declare that this thesis entitled

APPLICATION OF SENTIMENT ANALYSIS IN E-COMMERCE RECOMMENDATION SYSTEMS

is the result of my own research work except as cited in the references.

This thesis has not been accepted for any degree and it is not concurrently submitted in candidature of any other degree."

Name of Supervisor:	**ONG WEI AUN**
Signature:	……
Date:	24 April 2022

ACKNOWLEDGEMENT

Being able to complete this project brings me joy and a lot of knowledge and experience. First and foremost, I would like to thank my supervisor, Mr Raheem Mafas for his continuous guidance and support throughout the project. A special thanks also go to my second supervisor, Prof. Dr R. Logeswaran for his valuable insights, especially for the project report and presentation.

Besides, I would also like to thank all the lecturers at Asia Pacific University of Technology and Innovation (APU) for sharing their knowledge and equipping me with the knowledge needed to complete this project successfully. Lastly, I thank my family and friends for their continuous encouragement along the way to the completion of the project.

ABSTRACT

As technology advances at a rapid pace, artificial intelligence nowadays is playing an important role in digital marketing. With the increasing global adoption of e-commerce around the world, product reviews are playing an important role in customers buying decisions. Besides doing descriptive analysis that collects data like sales volume and product star ratings, product reviews also can reflect the performance of a single product in the market. However, product reviews left by customers may be fake or irrelevant to the products. Techniques of sentiment analysis (i.e opinion mining), can be used to extract data from these feedback reviews and filter spam reviews. Spam and irrelevant reviews online can mislead and force buyers to make purchases they would not have made otherwise. Therefore, this capstone project presents an application of sentiment analysis in e-commerce recommendation systems. The dataset used in this study is retrieved from an online data source made available by Amazon. Data exploration and pre-processing were done to explore more insights into the dataset. Next, natural language processing techniques are used to analyse the sentiment of the reviews and create a sentiment score. The sentiment scores are then used to create rankings, coupled with a search engine for users to find relevant products to their interest. A list of product recommendations is also provided with the search results. Overall, the project will study the development of a framework for e-commerce recommendation engines where the sentiment analysis of the product reviews can affect the recommendation score for the products.

TABLE OF CONTENTS

Contents

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Research Background

In this age of rapid advancement in technology, one of the booming sectors in the market is the e-commerce industry. Online retail purchases are growing every year and analysts are forecasting the market to grow by $10.87 trillion from 2021 to 2025, progressing at a CAGR of almost 29% for the forecast period (Infiniti Research, 2021). With that, the number of customers online and products available are also growing along with the industry boom. To improve the experience of shopping online, companies have implemented review systems on their shopping platforms to establish trust with customers. Reviews are given by customers to give feedback to companies and other users by recalling their own experience using the services or products. On the other hand, by reading other users' comments, one can gauge the performance of products they are about to purchase. This is because the information inside the comments can influence the customers' purchase decisions. Despite the benefits of product reviews, the immense increase in product reviews also resulted in comment information overload, making it difficult to gather meaningful information in a short time. Other than that, fake reviews are also a growing problem in the e-commerce industry. Customers can be incentivized to give good reviews to help companies grow their businesses (Laura Hautala, 2021).

Sentiment analysis is a text classification technique that focuses on subjective statements. It can process opinions to learn about perception. Also known as opinion mining, it collects and examines opinion or sentiment words using natural language processing (NLP) techniques to identify the sentiments from users about a specific subject and its characteristics (Sasikala & Mary Immaculate Sheela, 2020). People want to receive advice from others to ensure they make purchases wisely, which is why opinion mining is so popular. In the discipline of data mining and natural language processing, determining subjective sentiments in large amounts of social data is important to understand the opinion of the public. To make profitable business decisions, sellers also want to know which elements of their products are more popular among the consumer population. Opinion content may be found in abundance online in the form of blogs, forums, social media, and review websites, among other places. These contents are expanding every single day, with more and more public-contributed content being added

regularly. Millions of reviews should be analysed and aggregated to make a quick and efficient judgement, which is beyond the control of manual procedures. Therefore, with the help of sentiment analysis algorithms, the process can be automated with little or no user interaction.

In traditional shopping methods, products are usually mass-produced for a particular market and target audience. However, while online shopping became popular, competitive markets have to provide different products and services to different users with different needs and wants. Online platforms allow retailers to customize the relevant products to the users through recommendation engines or recommender systems. A recommendation engine, also known as a recommendation system, is an algorithm that works to recommend relevant products to consumers so that they can have a better user experience (Karnan & Seenuvasan, 2017). A good recommendation engine can produce a large amount of revenue for some industries. Therefore, applying sentiment analysis to such systems can help to make better recommendations to consumers and in turn benefits the sellers. Among the industries that benefitted from recommender systems is the restaurants industry. Asani et al. proposed a recommendation framework that identifies the food preferences of individuals from their comments and in turn provides these insights to the restaurants. The authors used a semantic approach to cluster the name of foods extracted from the comments and analysed their sentiments about the food. In the end, nearby open restaurants are recommended based on their similarity to user preferences. The proposed system is measured using precision, recall and f-measure metrics. It is found that the recommendations provided have a precision of 92.8%, which gives users a high degree of precision (Asani et al., 2021).

1.2 Problem Statement

Existing recommendation engines in e-commerce websites usually generate personalized user recommendations based on their preferences like recently-purchased items and item categories. However, in the case where a user is finding something new, the recommendation engines may recommend products based on sales, popularity, or product ratings from other users (Sivapalan et al., 2014). Since Shopee incentivize users for leaving a review (Shopee Malaysia, 2021a), users tend to leave irrelevant reviews not related to the purchases they made.

Spam and irrelevant reviews online can mislead and force buyers to make purchases they would not have made otherwise. Amazon also banned incentivized fake reviews back in 2016 to protect its sellers and buyers (Amazon, 2016b). Those reviews deprive potential purchasers of a fair and honest evaluation of products. As a result, buyers may be disappointed with the goods they bought. To improve on this situation, Shopee also removes spam product reviews that contain content that does not contribute to the perspective of the product(Shopee Malaysia, 2021b). However, if the company do not remove the spam reviews in time, it may affect customers' purchase decisions.

In light of this, this study intends to implement sentiment analysis in the framework of recommendation systems. By including the sentiment scoring of product reviews as quantification of explicit feedback, it can combine with other implicit feedback like clicks or views in the recommendation engine to help recommend better products to the customers.

1.3 Aim of the Study

The main aim of the research is to develop a framework for recommendation engine with weightage of customer reviews and product rating classification. This can be done by performing sentiment analysis on the customer reviews, classifying them and combining them in the recommendation mechanism of recommender systems.

1.4 Objectives of the Study

This study has several objectives to be achieved:

1. To identify the factors that influence customers' online shopping experience
2. To investigate the correlation between the product rating score and the reviews given by customers
3. To rank products based on the combination of reviews sentiment scoring and star ratings

1.5 Research Questions

The following questions will be addressed in this study:

1. What are the factors that influence customers' online shopping experience?
2. How close is the correlation between the product rating score and the product reviews?
3. Can we make better product recommendations to shoppers by including product reviews as a part of the recommendation systems?

1.6 Scope of the Study

The scope of the research will be limited to developing a framework for recommendation engine with weightage of customer reviews and product rating classification. The data to be used in the study should include both product ratings and product reviews by different users. This study intends to use natural language processing techniques to perform sentiment analysis on customer reviews. By using the sentiment scoring to combine with the product ratings in the recommendation mechanism of recommender systems, the study emphasizes creating product rankings based on their combined scores. This will be more related to non-personalized recommendations since personalized recommendations will require more data like past purchased items and purchasing behaviours like adding products to checkout carts. However, some irrelevant reviews are in the form of photos or videos. With that, it is highly uncertain

that this study will be useful enough and further studies will have to be done with image recognition techniques.

1.7 Significance of the Study

The implementation of recommendation systems plays a vital role in e-commerce companies where they aim to recommend relevant and good products to their customers. In addition to this, the companies rely on their customers' feedback to constantly improve their system performance. With a large pool of users in the community providing reviews in the system, it is important to understand user satisfaction during their purchases. It will be useless if the system upgrading is not aligned with user satisfaction. Therefore, this study will prove to be vital to the e-commerce industry in their recommendation systems to make better product recommendations to the users and help them make better purchase decisions. Only when the user satisfaction in the shopping environment is good, then it can improve the revenues of the companies and also their associate sellers.

1.8 Structure of the Report

With the aim of developing a framework for recommendation engine with weightage of customer reviews and product rating classification, the project report is organized into 6 chapters as follows:

Chapter 1 is the introduction chapter where an overview of the background of the study is presented to provide the idea of the study. The problem statement, aim, objectives, scope and significance related to the study are also identified and briefly described in this chapter.

Chapter 2 consists of a detailed review of past relevant literature. The applications and methods in sentiment analysis and recommendation systems are analysed critically to provide theoretical analysis on the application of sentiment analysis in recommendation systems.

Chapter 3 focuses mainly on the research methodology where the research approach in the study is highlighted. The dataset used in this study will be explored and the details of the data pre-processing methods are described. The techniques used in finding the sentiment scoring of the reviews are also discussed in this chapter. With this, the Python language will be used in the Google Colaboratory environment.

Chapter 4 outlines in detail the steps taken and techniques applied in the data preparation, sentiment analysis and recommendation system implementation.

Chapter 5 will discuss the results obtained after running the sentiment analysis and the recommendation system. Lastly, Chapter 6 will conclude the project by summarizing the processes in the project and emphasizing the contributions of this study.

1.9 Research Plan

This study was conducted in phases where the approximate duration is around 6 months, starting from 20 September 2021 to 24 April 2022. The first month was used to find suitable data from online databases based on the relevance to the project. After finding a suitable dataset, the dataset is explored, and data cleaning is done in the following 2 months. The first draft of the report consisting of the introduction, literature review and methodology was also done within that 2 months. In the 4th month, natural language processing was done by using NLP techniques and sentiment scores are created for the text reviews. Finally, in the 5th and 6th months, a search engine and a recommender system are built for the recommendation purpose. Report writing was done simultaneously with the data processing stages and is scheduled to be submitted by end of April 2022.

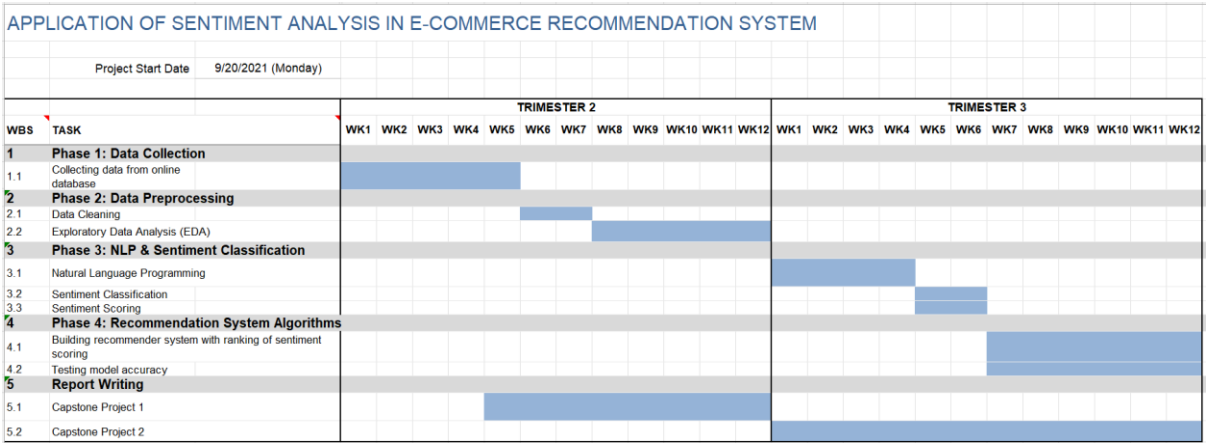

Figure 1.1: Gantt Chart for Research Plan. The study is planned to be done in 2 trimesters, which are from Sep 2021 to Apr 2022. It is mainly separated into 4 phases: data collection, data preprocessing, sentiment scoring and recommendation system.

CHAPTER 2

LITERATURE REVIEW

2.1 Sentiment Analysis

Sentiment analysis can be done to analyse individuals' emotions, expressions, attitudes, excitement, opinions, and viewpoints toward certain entities. The opinion sentiment can be made up of five components as follows: {e; t; s; h; dt}, where e denotes the entity, t denotes the target aspect or feature for the sentiment, s denotes the sentiment of the target, h denotes the opinion holder, and dt is the timestamp of the opinion. There are 3 main different levels where sentiments can be analysed, which are the document, sentence, and aspect levels.

Document-level sentiment analysis seeks to determine whether a document as a whole reflects a negative or positive mood or neutral opinion (Alqaryouti et al., 2020). Every single document is considered as one information unit and categorized depending on the opinion holder's overall sentiment towards a single entity. Therefore, document-level analysis is only effective when the document is written by one single individual and can be not effective when evaluating different entities. Many approaches to document-level sentiment analysis have been proposed. Duyu T. introduced a sentiment-specific representation learning framework for document-level sentiment analysis. The semantics of the document is first broken down to learn the representations of sentences through word embeddings and sentence structure and composition. After that, the document representations are calculated through document composition based on sentence representations and discourse analysis. Finally, the learned document representations are applied as features and run with a document-level sentiment classifier with Support Vector Machine (SVM) (Tang, 2015).

Similar to the document level, sentence-level treats sentences like short documents. The primary purpose of sentence-level sentiment analysis is to figure out whether the language represents a good, negative, or neutral viewpoint (Birjali et al., 2021). Kupivalova et al. introduced an architectural design for a semantic search tool by using natural language processing. The design first performs lexical analysis and parsing to identify and analyse the structure of words for grammatical errors. After that, it goes to semantic analysis where it maps syntactic structures and objects within the task area and finds the meaning of an immediate success sentence by discourse integration. Lastly, it performs pragmatic analysis to re-interpret

what is the actual meaning of the sentence (Kupiyalova et al., 2020). However, to properly perform sentence-level sentiment analysis, the statement must be defined as objective, defining facts, or subjective, expressing feelings and ideas. In this case, an opinion may seem positive relating to the entity but not all the aspects of the target may be satisfied – which leads us to the aspect-level of sentiment analysis.

Aspect-level performs more fine-grained analysis because it identifies what people like or dislikes according to specific aspects of entities, where it can help to detect sentiments. It emphasizes the features of entities (eg: product aspects) rather than the sentiment of paragraphs or sentences (Alqaryouti et al., 2020). It is done by breaking down the entity into separate aspects and classifying each aspect into the sentiment of positive, negative, or neutral, which is known as the aspect sentiment classification. Finally, the results of the sentiment will be summarized to identify the sentiment regarding the entity. This type of analysis is applied in various real-life situations. For example, through aspect-level analysis, companies can identify which features or parts of a product are appealing to customers in the effort to improve their products. An approach to feature-based sentiment analysis was proposed by Cai and Liu, who used the Apriori algorithm to expand the sentiment ambiguous lexicon, which are triples of sentiment object, sentiment word and sentiment polarity. The authors use association rule mining of sentiment ambiguous words collocation set and find out the element relationship between the words. However, complex products phones and laptops have many features or aspects which need more in-depth study because the computer parts contain more opinion words that are associated directly with the target domain, while by passing its feature, a lot of valuable information can be missed (Cai et al., 2017).

To reach better performance, two or more levels of sentiment analysis can be combined instead of using only one single level. Mai and Le proposed a combined approach of both sentence and aspect-level sentiment analysis of product comments on YouTube (Mai & Le, 2021). It is assumed that there is a strong relationship between sentence-level and aspect-level sentiment analysis. This is because the polarity of sentiment on sentence-level and aspect-level can affect one another and therefore the joint approach can solve the problem of the two levels together. After preprocessing is done on the comments, a BERT based model (Devlin et al., 2019) was applied to identify the sentiment on both aspect and sentence levels. Finally, the analysis results are compiled to produce statistical reports on the target product.

Another type of sentiment analysis can be considered as the concept-level. Unlike the word-based approaches like aspect-level or sentence-level, concept-level uses web ontologies and semantic networks to perform semantic analysis of text, allowing the combination of conceptual and affective words associated with natural language. For example, if the concept of cloud computing is analysed through a word-based approach, the "cloud" word would be wrongly associated with the weather instead of virtualization. Nevertheless, despite the benefits of concept-level sentiment analysis, it can be constrained to the limits of the knowledge base. Other than that, it also cannot detect structural information in opinions to effectively detects the polarity expressed by natural language like negation words. (Poria et al., 2014)

When negations are processed inappropriately, it can lead to biases and misclassification of sentiments. Identifying negation presence can improve the accuracy of sentiment classification by identifying the parts of sentences that would have different polarities after pairing up with the negation term. Mukherjee et al. proposed a sentiment analysis approach to include negation identification and negation scope marking for negation handling. They proposed to detect explicit negation through a customized negation marking algorithm and perform sentiment analysis experiments with multiple machine learning methods on Amazon reviews of cell phones. (Mukherjee et al., 2021).

Other than that, in some data pre-processing methods, negation words are removed because they are in Stop-word lists or are implicitly ignored as they appeared as neutral sentiment in a lexicon which does not affect the final polarity. Therefore, inverting the polarity does not make this work easier because negation words may appear in a sentence without affecting the sentiment of the sentence. A common approach for negation handling is tagging the words with 'NEG_' after negation until the first punctuation. However, this may make the tag un-negated, and this method also cannot handle negation and conjunction in one sentence. Amalia et.al suggested using a rule-based method to determine which words were negated by applying the Indonesian language's negation syntactic rules to figure out what the scope of negation is. By using syntactic rules and tagging 'NEG_' with SVM classifier with RBF kernel, it can improve negation handling with an improvement of 3 to 5% against existing negation handling. (Amalia et al., 2018)

2.2 Spam Detection in Sentiment Analysis

In the domain of sentiment analysis, spam detection is critical. Fake reviews and spam can harm brands' reputations and artificially affect consumers' perception of products, services, corporations, and other entities since online opinions can influence consumer buying decisions. (Silva et al., 2018)

Driven by profit intentions, spammers can give fake reviews about the target store, hence deceiving the customers about the truth. To tackle this issue, Peng & Zhong introduced a framework for spam review detection. In the study, the authors integrate the techniques of sentiment analysis into spam review detection. By using a shallow dependency parser, the sentiment score is computed from the natural language text. Next, a series of discriminative rules are created through intuitive observation. In the end, they established a time series combined with discriminative rules to detect the spam reviews efficiently. (Peng & Zhong, 2014)

Other than that, Kauffman et al. also introduced a Fake Review Detection Framework (FRDF) which detects and removes fake reviews using natural language processing techniques. The framework automatically analyse product reviews and transform them into negative and positive user opinions in a quantitative score (Kauffmann et al., 2019). It was tested on high-tech industries' product reviews and the brands were rated according to consumer sentiment. It is found that brand managers and consumers find this tool helpful in decision making alongside the 5-star scoring system.

Kontsewaya et al. also reviewed several machine learning algorithms in the study of spam detection (Kontsewaya et al., 2021). They found that logistic regression and Naïve Bayes give the highest level of accuracy – up to 99%. They also suggested creating a more intelligent spam detection classifier by combining filtering methods or algorithms. Besides, Naïve Bayes algorithm also performs the best as the base classifier at an accuracy of 93% in a study of the effectiveness of semi-supervised learning approaches for opinion spam classification. (Ligthart et al., 2021)

2.3 Recommender Systems

Among the most popular data mining methods applied in the e-commerce industry is the association rule. Association rules find the relationship between different items transacted in different transactions. When two products are purchased together, the relationship between the two items can be identified, which can be very useful during recommendations to new users looking to make new purchases. However, there is a problem with association rules. Association rules take a lot of data and time to perform hence making it not effective when it comes to database scaling as transactions are building up in real-time day by day. (Sivapalan et al., 2014)

With that, the working principle behind a lot of recommender engines is based on two fundamental approaches, which are content-based filtering and collaborative filtering. (Felfernig et al., 2019) Originated from the idea of word-of-mouth promotion, collaborative filtering is based on customers' activities, interests and behaviours and predicting what they will like based on their similarities with customers. Collaborative filtering can be achieved on a heuristic-based, model-based or hybrid one combining both methods. The heuristic-based method takes in data and utilizes k-nearest neighbour classification to identify the similarity between users or items. On the other hand, the model-based method uses training data to build a model before validating it through training data. While the heuristics-based model considers the entire database and the customers to create recommendations for the new customer, the model-based approach only relies on the past existing customer information as the model building test data. The biggest problem in Collaborative Filtering is the sparseness of observed values. It means feedback is observed in a very small portion of all possible user-item pairs. However, the Matrix Factorization model is known to work better than other models even if the data is sparse. (Karnan & Seenuvasan, 2017). Besides the implicit feedback that is generated by users, there is also explicit feedback from users like ratings or votes. Liu et al. performed matrix factorization models that are trained from both implicit and explicit responses, and they discovered that the algorithm can improve recommendation quality (Liu et al., 2010).

Next, content-based filtering is based on the things liked by customers and the keywords for the item. Therefore, the algorithm will identify and recommend similar products to those previously purchased by the customers. Hybrid recommendation systems are where both

recommendation approaches are combined to compensate for the disadvantages of each approach. Some lesser-known recommender systems also include utility-based recommendation and group recommender systems. Utility-based recommendation recommends products based on multi-attribute utility theory where items are evaluated on knowledge-based evaluations like the attributes of the items. Meanwhile, group recommender systems recommend products based on group decision heuristics instead of individual personal preferences.

Other than that, Shen et.al suggested a novel algorithm known as Sentiment Based Matrix Factorization with reliability (SBMF+R) to utilize reviews to provide reliable suggestions. (Shen et al., 2019) The algorithm is made up of three parts where a sentiment dictionary is first constructed to change the feedback reviews into sentiment scores. Next, user reliability measures are designed to combine user consistency and reviews. Following that step is where the rating, reviews and feedback are incorporated into a probabilistic matrix factorization framework to enhance the recommendation system performance. The framework managed to perform better than the other commonly used algorithms.

Stochastic learning algorithm can also be used in recommendation engines to classify the products as positive and negative where positive products based on ratings and reviews are shown in the recommendation list. (Karnan & Seenuvasan, 2017) It is found in the study that stochastic learning algorithm outperform collaborative filtering in terms of precision and recall performance.

2.4 Summary

In a nutshell, Chapter 2 compiles past literature about the techniques of sentiment analysis and its application in recommendation systems, especially in the e-commerce domain. The previous works of scholars have shown that sentiment analysis can be done at a different level of depth. Other than that, spam detection has also extensively been applied in the e-commerce industry where it can help in decision making apart from the 5-star rating. Moreover, the types of recommendation systems and their working mechanisms were elaborated on and discussed to review their applications in creating recommendations. Recommendation systems have been widely applied in e-commerce already, however they are mainly content-based and collaborative-based which are mainly based on implicit feedback like clicks or purchases. The

goal of this study will be to unify both explicit (reviews, ratings) and implicit feedback to generate recommendations.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter presents the research approach adopted for this study. The entire research methodology outline is provided and described in this chapter. The stages in the methodology including the data collection, data pre-processing, data exploration and natural language processing techniques are discussed in this chapter in detail.

3.2 Research Approach

The schematic representation of the overall research methodology proposed in this study is shown in Figure 3.1. There are five main processes in this research framework, which are data collection, data pre-processing, exploratory data analysis, sentiment analysis and recommendation system. All of the processes and the methodologies applied are explained in the following sections.

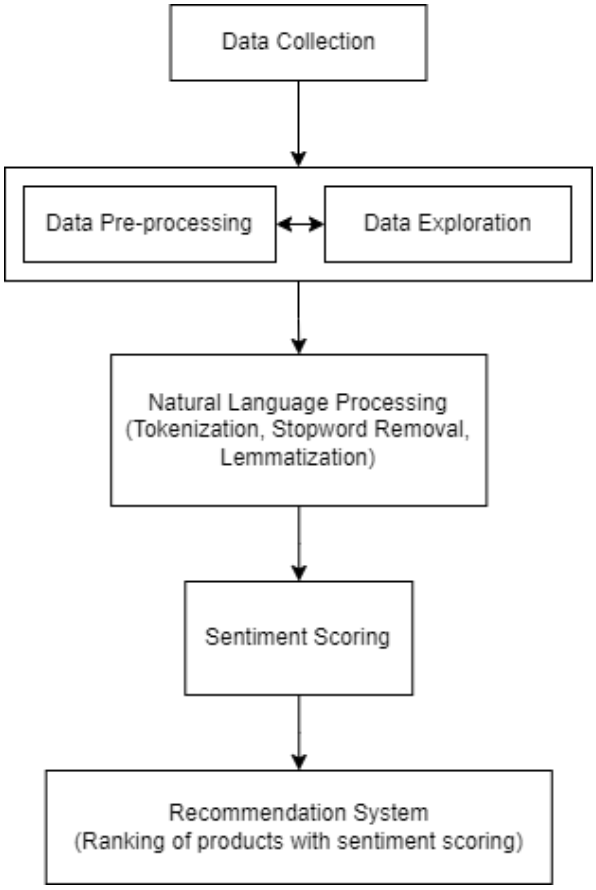

Figure 3.1: Research Methodology. There are 4 main phases in the research methodology, which are data collection, data pre-processing, NLP & sentiment scoring and recommendation system.

3.2.1 Data Collection

The first stage of the research is the data collection process. The dataset for this study is sourced through a Node.js module used to crawl product reviews from Amazon. (Niemi, 2015). It is a collection of reviews written in the Amazon.com marketplace and associated metadata from 1995 until 2015. There are multiple Amazon categories available and the chosen one is from the outdoors category. The dataset that is collected for the study, comprises the metadata of outdoor products and their associated reviews. Every row in the datasets represents the details of every review.

3.2.2 Description of Dataset

The dataset to be used in this study have over 2 million observations collectively and 15 variables describing the characteristics of the reviews. Table 3.1 describes the variables present in the data set used in this study.

Table 3.1: Description of Variables in Dataset

#	Attribute	Description
1	Marketplace	2 letter country code of the marketplace.
2	Customer_id	The unique ID of the customer.
3	Review_id	The unique ID of the review.
4	Product_id	The unique Product ID the review pertains to.
5	Product_parent	Random identifier that can be used to aggregate reviews for the same product.
6	Product_title	Title of the product.
7	Product_category	Broad product category that can be used to group reviews.
8	Star_rating	The 1-5 star rating of the review.
9	Helpful_votes	Number of helpful votes.
10	Total_votes	Number of total votes the review received.
11	Vine	The review was written as part of the Vine program.
12	Verified_purchase	The review is on a verified purchase.
13	Review_headline	The title of the review.
14	Review_body	The review text.
15	Review_date	The date the review was written.

3.2.3 Data Pre-processing

After collecting the data, data pre-processing is important to make sure the data suits the requirements of the study. The tasks in data pre-processing include data cleaning, data transformation and data reduction etc. All data exploration, pre-processing and coding in this study will be done in Google Colaboratory.

First of all, the data collected will be analysed to identify the most suitable variables to be used for the sentiment analysis process. Some of the variables like the verified purchase, vine and marketplace may only be used for data exploration to better understand the data. Meanwhile, variables like star ratings and review body will be used as the main input for the sentiment analysis and the rankings.

Data cleaning process will be done to identify any missing part of the data before modifying, deleting or replacing them according to necessity. Missing values in the dataset have to be treated properly. If the total missing values are equal to less than 5% of the total data, they will be ignored and removed from the dataset (Salgado et al., 2016). However, if a substantial number of missing values is found in the dataset, data imputation methods like the mode or mean imputation will be applied to fill in the missing values.

Unrelated variables like marketplace and product category will also be excluded from the further parts of the study because they can be irrelevant to the recommendation of products as the dataset are from the US market and are all outdoor products. Since sentiment analysis of the text reviews is the most crucial part of the study, it will be studied in deeper detail in the following steps of the methodology.

3.2.4 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is done to investigate the data to find out patterns or irregularities and check assumptions with the use of visual and graphical representations (Jebb et al., 2017). This task will produce some visual outputs like piecharts and histograms to provide a better understanding to decide on further data pre-processing if deemed necessary. Both data pre-processing and the EDA process will be conducted concurrently to ensure all issues discovered in the dataset are resolved.

3.2.5 Sentiment Analysis

Natural language processing (NLP) techniques can be used to classify the textual reviews into positive or negative sentiment and calculate a sentiment score (Fang & Zhan, 2015). Product reviews consist of subjective personal opinions on the products and they all contain some sentiment sentences. To perform sentiment analysis, data preprocessing like tokenization, removing stopwords and lemmatization (or stemming) must be done.

3.2.5.1 Tokenization

The first step in sentiment analysis is tokenization, which is breaking down every sentence into individual letters, words, or other substring sections of equal or unequal length. By breaking down the sentences into words, the connection between words can be broken down as well. Individual words make it more convenient and efficient to analyse the text data by examining the words appearing in an article and the number of word appearances to give insights into the sentiment (Ofer et al., 2021). After the tokenization process, every sentence will transform into a list of words, symbols, digits, and punctuation. The symbols, digits and punctuation can then be removed because they do not contribute information to the sentiment analysis.

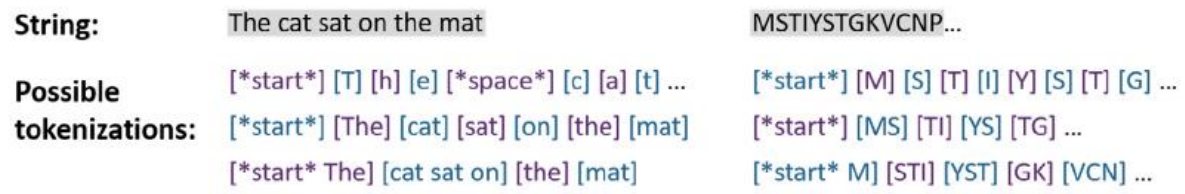

Figure 3.2: Tokenization (Ofer et al., 2021). Tokenization is breaking down every sentence into individual letters, words, or other substring sections of equal or unequal length.

3.2.5.2 Stopwords Removal

After transformation, though the sentences are much cleaner, there will still be some words like “and”, “I”, “we” etc. that are useless. They are also known as stop words. These words will not affect the understanding of sentences even if being removed. Therefore, the stopwords will be removed by importing stopwords from the NLTK library (Ramachandran & Parvathi, 2019).

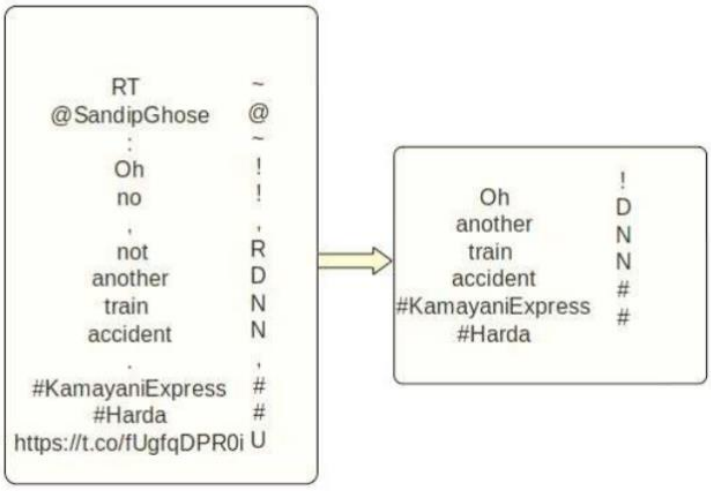

Figure 3.3: Stopwords Removal (Ramachandran & Parvathi, 2019). These words can be removed from sentences without affecting the understanding of sentences even after being removed.

3.2.5.3 Lemmatization

After tokenization and removing stopwords, the sentences will be transformed into a list of meaningful words. This can be done through stemming or lemmatization. Stemming is a process that removes the suffixes or prefixes used with a word while lemmatization, unlike stemming, reduces the words to the base form (Ramachandran & Parvathi, 2019). In light of that, it is important to remove grammar tense and transform each word into its base form so that the number of the appearance of each word can be counted. Therefore, lemmatization is a vital process in text transformation.

3.2.5.4 Valence Aware Dictionary for Sentiment Reasoning (VADER)

As an alternative to the data pre-processing methods mentioned above, VADER is another tool that can use raw text to analyse the sentiment without pre-processing. Therefore, VADER from the NLTK library will also be used to identify how positive or negative a review is, through a scale of -1 to 1 where -1 indicates the most negative sentiment and 1 is the most positive with 0 as neutral sentiment (Hutto, C.J. and Gilbert, 2014).

3.2.6 Recommendation System

The sentiment scores will be combined with star ratings as a ranking to be implemented as part of the recommendation mechanism in this study. Since metrics like sales volume and popularity involves getting data like prices and the number of products sold, they will not be

included as part of the study because the dataset used in the study does not have these data. With that, the recommendation system to be developed in the study will only be using the combined scores and title relevancy as part of the recommendation system.

3.3 Summary

This chapter described the research methodology used for the study. There are five main stages in the methodology, which are data collection, data pre-processing, data exploration, sentiment analysis and recommendation system. The theoretical ideas behind every process were also briefly explained in this chapter. Idealistically, a recommendation system would be developed and tested in this study by using the star rating and sentiment scoring from the product reviews. The techniques and specific details of the implementation will be further discussed in the following chapter.

CHAPTER 4

IMPLEMENTATION

4.1 Introduction

This chapter mainly focuses on the implementation of sentiment analysis and the recommendation system with a detailed explanation of the techniques used in this project. The dataset used in the project and the pre-processing methods will be described in detail. Some data pre-processing techniques like data exploration and missing data handling were applied during the data preparation stage. A detailed explanation for each step of pre-processing is provided. After data pre-processing is done, sentiment analysis will be done by using the text of the product reviews. Natural language techniques like tokenization, stopword removal and lemmatization will be applied for the sentiment analysis process. Finally, a recommendation system with a search engine will be created to help users find relevant products according to the sentiment scoring rankings.

4.2 Data Pre-processing

Data pre-processing is an important step to prepare the data ready for the requirements of the project. During initial exploration, the dataset was discovered to have some inconsistencies in data. Therefore, pre-processing techniques were used to prepare the data for the sentiment analysis process.

4.2.1 Initial Data Exploration and Dropping Variables

The dataset to be used in this study have over 2 million observations collectively and 15 variables describing the characteristics of the reviews. Table 3.1 in the previous chapter describes the variables present in the data set used in this study. However, upon looking into the details of the variables, it is found that some irrelevant variables can be dropped from the dataset.

The ‘marketplace’ variable consists of only one class, which is the US. This is because Amazon has different marketplaces for customers in different regions of the world and this dataset comprises only products in the US.

Other than that, the ‘*product_category*’ variable also comprises only one class, which is outdoors. This is because the dataset retrieved is of the outdoors category, hence making the

variable unnecessary for the upcoming data processes. The '*product_parent*' variable is the identifier that can be used to aggregate reviews for the same product.

There are 2 columns related to votings from other users toward a particular review, which are '*helpful_votes*' and '*total_votes*'. However, this voting method will not be used to contribute to the recommendation mechanism used in this study. Therefore, both variables will be dropped as well. Last but not least, the '*review_headline*' contains the title for the reviews, which may only be the summary of the review body. Therefore, sentiment analysis can directly be done to the review body instead of the review title for a better representation of the sentiment in the reviews.

In a nutshell, these 6 features (*marketplace, product_parent, product_category, helpful_votes, total_votes, review_headline*) in the dataset will be dropped before running exploratory data analysis.

```
# Dropping variables
df2 = df.drop(['marketplace','product_parent','product_category','helpful_votes','total_votes','review_headline'], axis = 1)
```

Figure 4.1: Dropping variables. 6 different variables are dropped from the dataset before running exploratory data analysis.

4.2.2 Missing Value Treatment

Data preparation is the most important task in the data science methodology which can strongly affect the outcome of the research. One of the main tasks during data preparation is handling the missing values in the dataset. To properly handle missing data, the proportion and patterns for missing data must be identified first before deciding on the imputation method (Salgado et al., 2016). There are 3 main types of missing values, which are missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

```
df2.isnull().sum()
```

```
customer_id            0
review_id              0
product_id             0
product_title          0
star_rating            3
vine                   3
verified_purchase      3
review_body          135
review_date           13
dtype: int64
```

Figure 4.2: Missing values in the dataset

As shown in Figure 4.2, it can be observed that the missing values are considered MCAR where there is no hidden mechanism related to any features and the tendency for a data point to be missing is completely random. With the large dataset of around 2.3 million observations, it is assumed that the remaining subsample of non-missing data is representative of the population and will thus not bias the analysis towards the missing subgroup. Therefore, the listwise deletion method is done to remove all the missing values in the dataset. This is because of the simplicity of the method, and it is reasonable to use it when the number of dropped observations is relatively small compared to the total.

```
#Drop Null values in the columns in Pandas
df2=df2.dropna()
```

```
df2.isnull().sum()
```

```
customer_id          0
review_id            0
product_id           0
product_title        0
star_rating          0
vine                 0
verified_purchase    0
review_body          0
review_date          0
dtype: int64
```

Figure 4.3: Dropping the missing values. There are no more missing values after dropping the empty observations

4.3 Exploratory Data Analysis

Running exploratory data analysis (EDA) is an important process where the data distribution can be examined to maximize insight into the database and outlier identification (Komorowski et al., 2016). In this process, the remaining variables are explored by using bar charts and pie charts to understand the distribution of each variable.

4.3.1 Sales Seasonality

Retail sales seasonality refers to the regular fluctuations in sales throughout the year. It is important for e-commerce companies to monitor the seasonality of their sales so that they can expect to lose money for certain periods and adjust to balance out the profitable periods. Therefore, in this study, sales seasonality is also explored in this section through the dates from the reviews.

```
df2["review_date"] = df2["review_date"].astype("datetime64")
date_df = df2 [["review_date"]]
```

```
from matplotlib import pyplot as plt
%matplotlib inline
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")

plt.figure(figsize=(10,5))
sns.countplot(date_df["review_date"].dt.year, palette=sns.color_palette("pastel", 5))
plt.title("Distribution of Reviews through the years", fontweight='bold', fontsize=15)
plt.xlabel("Years")
plt.ylabel("Number of reviews")
plt.show();

print(date_df["review_date"].dt.year.value_counts().sort_index());
```

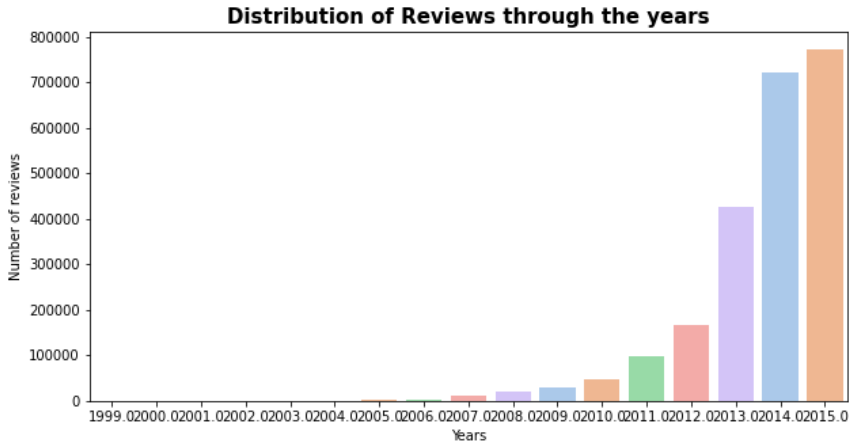

Figure 4.4: Distribution of Reviews through the years. The sales have been increasing steadily throughout the years from 1999 to 2015.

For the visualisation of the annual growth rate of the sales volume, the year part was extracted from the review dates and plotted with a bar chart as shown in Figure 4.4. It is observed that the sales have been increasing steadily throughout the years from 1999 to 2015. Furthermore, a jump in the number of reviews can be seen between 2012 and 2013, indicating the starting adoption of Amazon by the public.

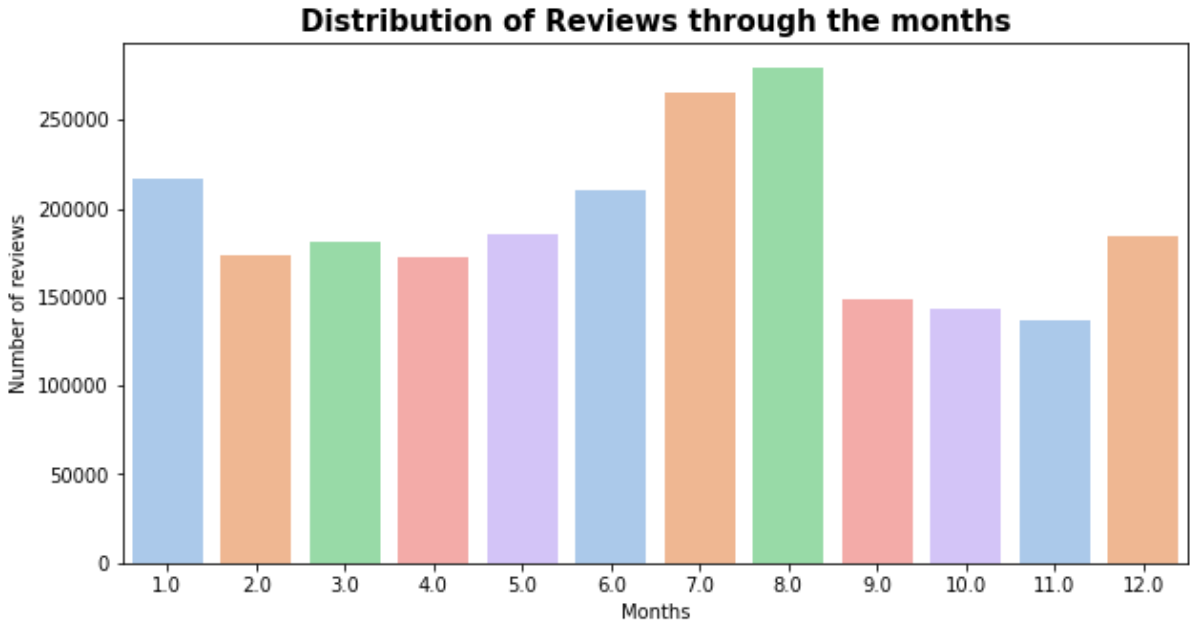

Figure 4.5: Distribution of Reviews through the months. The top 4 months are January, June, July and December.

Figure 4.5 shows the distribution of reviews through the months. The chart shows that the number of reviews peaks during the middle of the year. This can be explained by the weather seasonality when the summer is arriving. People start preparing for their outdoor activities and make their equipment purchases online. Besides that, January and December also show a higher than average volume of reviews. This period coincides with the shopping festive seasons like Black Friday, Christmas, Thanksgiving and Cyber Monday, hence leading to an increase in sales.

4.3.2 Amazon Vine Program

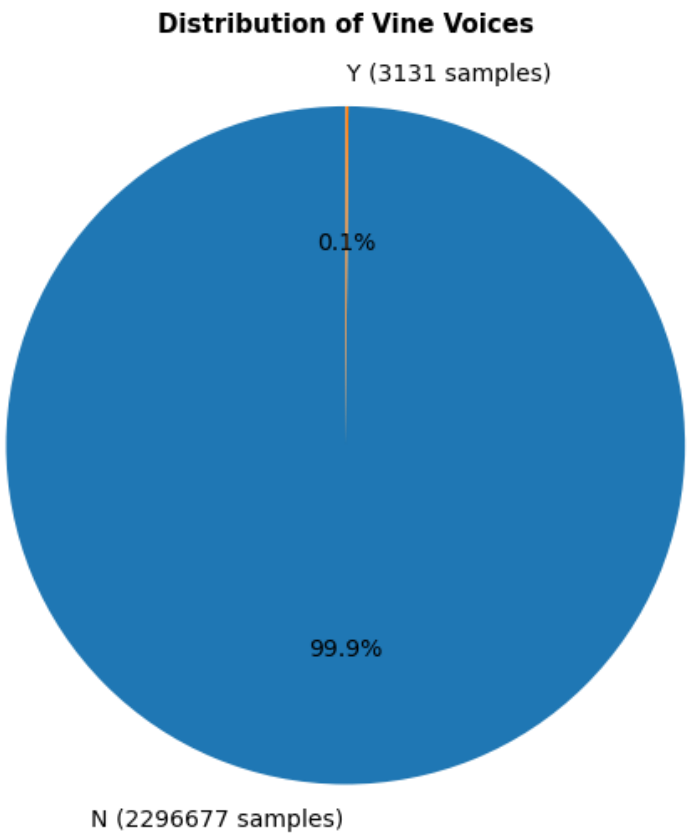

Figure 4.6: Distribution of Vine Voices. 99.9% of the reviews are non-Vine members because Vine Voices is an exclusive club of reviewers.

Figure 4.6 shows the distribution of the reviewers who are in the Vine program. Only a remarkably small percentage of the reviews (0.1%) are from Vine members. This is understandable when Amazon only invites the most trusted reviewers on Amazon to join as Vine Voices.

Amazon Vine is a program that invites reviewers to post their reviews on the Amazon platform. In return, the Vine members, also known as Vine Voices, get free products from participating vendors. These Voices are selected based on the helpfulness of reviews voted by other customers (Amazon, 2016a). Since this is an invitation-only program, only the most trusted reviews on Amazon are invited, which explains the skewness of the distribution in the dataset.

4.3.3 Verified Purchases

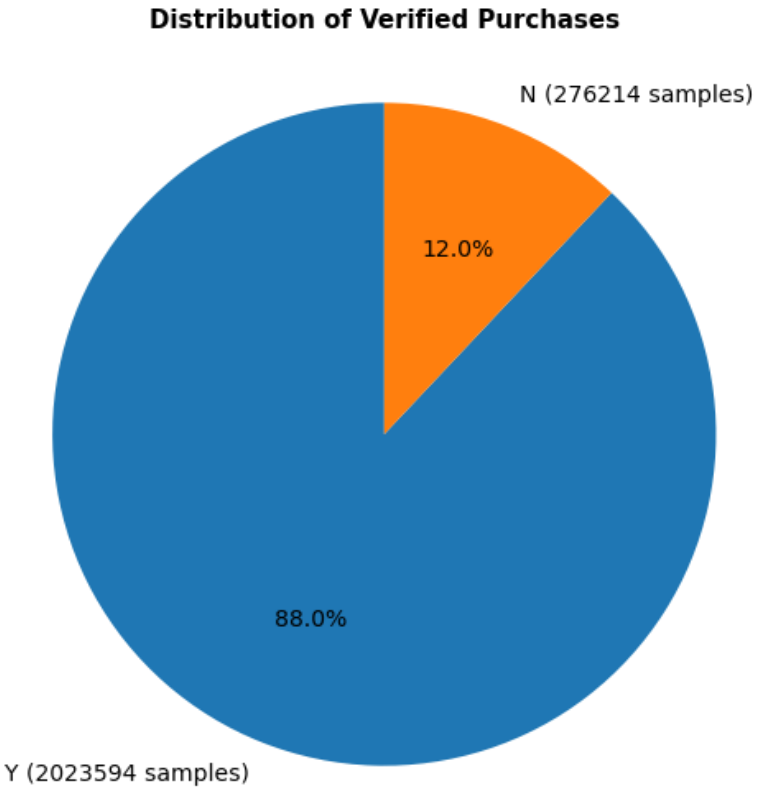

Figure 4.7: Distribution of Verified Purchases. 88.0% of the reviews have been verified as actual purchases.

The distribution of reviews that are verified as actual purchases is shown as a pie chart in Figure 4.7. 88.0% of the reviews have been verified as actual purchases. To reduce the possibility of having fake reviews, only verified purchases will be used for further processes of the study. The reviews with verified purchase status are filtered as shown in Figure 4.8.

```
#Drop verified purchase, filter out N values, left with Y values
df2 = df2[df2['verified_purchase'] != 'N']
```

Figure 4.8: Filtering only reviews with verified purchases status

4.3.4 Distribution of Ratings

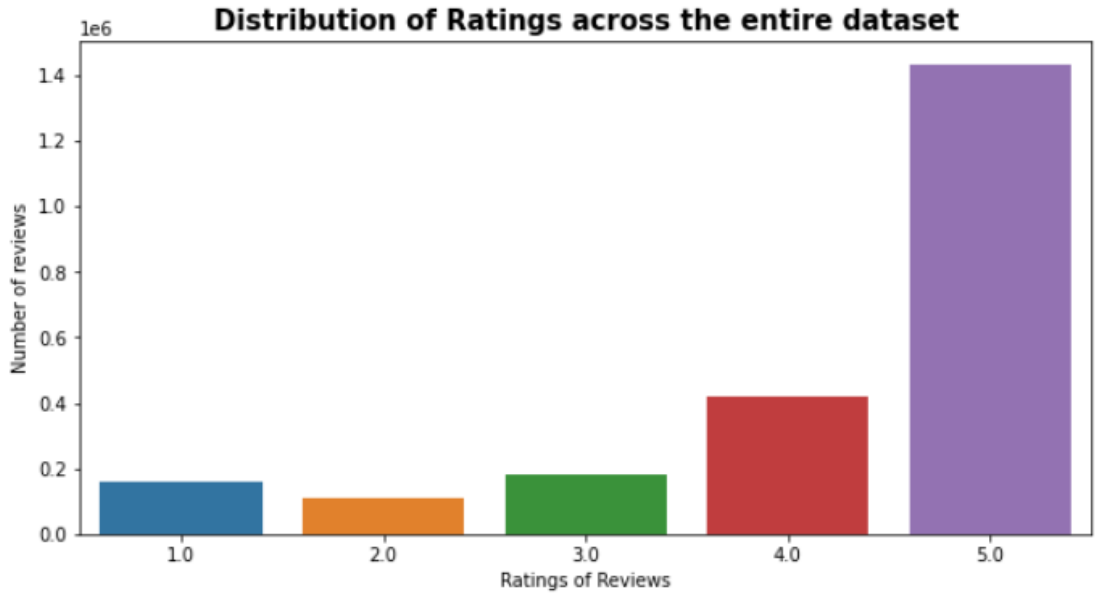

Figure 4.9: Bar Chart distribution of the Ratings. It is found that the ratings of 5 are in 62.3% of the dataset. A further 18.1% of the reviews have a rating score of 4.

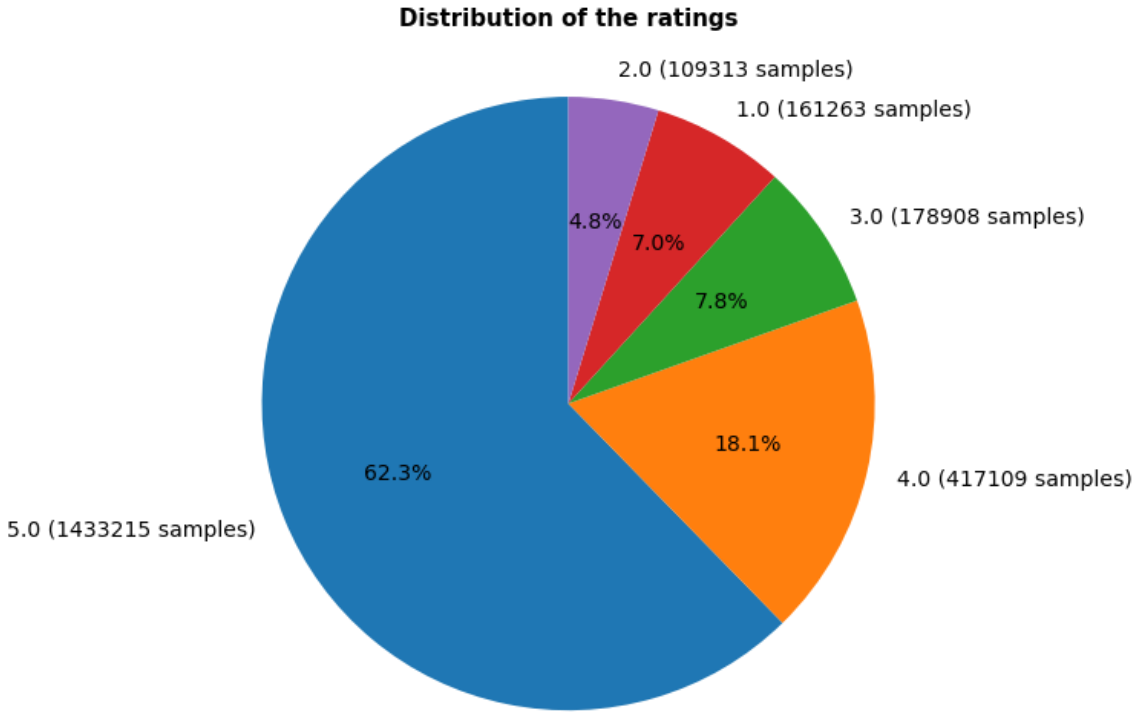

Figure 4.10: Pie Chart distribution of the Ratings

Figure 4.10 displays the pie chart of the distribution for the star ratings of the reviews. With a rating scale from 1 to 5, it is found that the ratings of 5 are in 62.3% of the dataset. A further

18.1% of the reviews have a rating score of 4. Therefore, the dataset is positively skewed as shown in Figure 4.9, with a combined 80.4% having ratings of either 4 or 5, indicating the overall satisfaction towards the products by the customers. However, the dataset is not edited despite the positive skewness. This is because these ratings reflect a measure of satisfaction by the customers, hence making it important for further analysis.

4.4 Sentiment Analysis

After running exploratory data analysis, sentiment analysis is done to extract the sentiment score from the reviews. There are multiple pre-processing techniques inside the process of sentiment analysis alone, such as tokenization, stop-words removal and lemmatization. For the sentiment scoring, VADER is chosen as the tool to create sentiment scores based on the pre-processed review text.

4.4.1 Data Reduction

As the dataset that is used in the study contains around 2.3 million reviews, the time taken to perform sentiment analysis is very long. With the need for text pre-processing like tokenization, stop-word removal and lemmatization, it is taking a toll on the processing power in Google Colaboratory. Thus, to make the data denser, the dataset is filtered where only the users who have given 50 or more ratings are included in the sentiment analysis process. Other than that, the sentiment scoring in the following steps will also be done in a simple way to save the processing time. The details are discussed in the following sub-chapter.

```
#Getting the new dataframe which contains users who has given 50 or more ratings

new_df=df2.groupby("customer_id").filter(lambda x:x['vader_sentiment_score'].count() >=50)
```

Figure 4.11: Filtering the users who have 50 or more ratings. This is done to make the data denser because the effort to process the data is both power and time-consuming.

4.4.2 VADER Sentiment Scoring

To perform sentiment scoring in this study, the VADER module in the NLTK library is used. This module is chosen because it does not need training data and can very well understand the sentiment of a text containing emoticons, slang, conjunctions, capital words, punctuations etc (Hutto, C.J. and Gilbert, 2014). Other than that, VADER does not need a lot of pre-processing where typical pre-processing techniques like tokenization, stop-words removal and lemmatization are not required. This is because VADER used a lexicon-based approach to determine the overall sentiment of the whole text.

The VADER algorithm can output sentiment scores into 4 different classes of sentiments, which are positive, negative, neutral, and compound. A compound score is used in this application for sentiment analysis, where the compound score is the sum of positive, negative & neutral scores which is then normalized between -1 (most extreme negative) and +1 (most extreme positive). Therefore, the higher the compound score indicates the higher the positivity of the text.

After obtaining the sentiment score through VADER, the new variable will act as a recommendation score to rank the products in the next stage of the project. The mean of the VADER scores grouped by product id is also calculated to find the relationship between the average VADER score and the actual star rating provided by the customers.

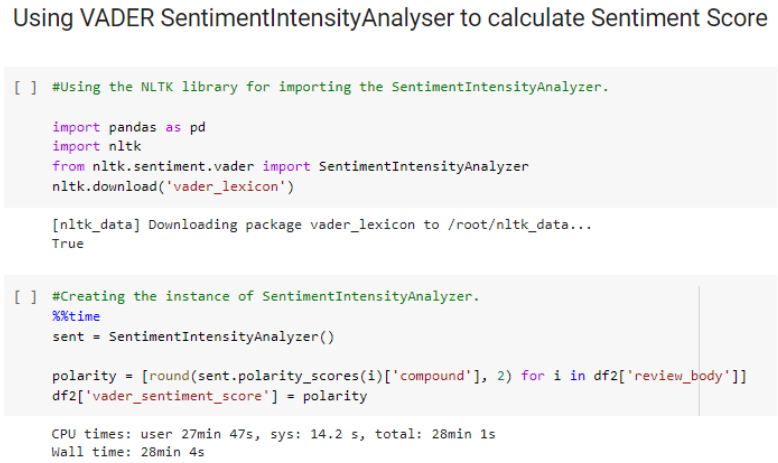

Figure 4.12: Code Snippet for VADER SentimentIntensityAnalyser

Furthermore, before combining both the actual star rating and the VADER score, the upper and lower limits of the sentiment scores are first rescaled from (-1 to +1) to (0 to 5) by using

the formula as shown in Figure 4.13 (Giannoulis, 2019) and the output is named as sentiment score. Next, both the actual star rating and the sentiment scores are added to generate the recommendation score of 0-10 with a weightage of 50% from both variables. The code implementation is shown in Figure 4.14.

$$Y = \left(\frac{X - X_{min}}{X_{range}}\right)n$$

Figure 4.13: Formula for rescaling upper and lower limits. (Giannoulis, 2019) It is used to rescale the VADER score (-1 to +1) to a scale of (0 to 5)

```
new_df['avg_sentiment_score'] = new_df.apply(lambda row: ((row['avg_vader_score']-(-1)) * 5/2), axis=1)
new_df['recommend_score'] = new_df['avg_rating']+new_df['avg_sentiment_score']
```

Figure 4.14: Code Implementation for VADER score rescaling and calculating the recommendation score. The average rating scores and the average sentiment scores are added to generate the recommendation score of 0-10 with a weightage of 50% from both variables.

4.5 Search Engine

After the sentiment scoring process, a search engine is created to help users find the relevant items from the dataset and users can do so by searching the keywords relevant to their buying intention. The keywords that are searched are matched with the words appearing in the product title of the products as the product title is an indication of the relevance to the products that the customers are finding.

The recommendation score helps the users measure how much an item is recommended. If there are no relevant results found from the search engine, it will return a print function telling the user to search again using other keywords. As the results are listed down, the results will be ranked according to the recommendation score.

```
print('Finding an outdoors product?\n')
searchTerm=input("Search for your outdoors product here (or enter 'exit' to exit):")

if searchTerm.lower() == "exit":
  print("\n----------------------------------------------\nThanks for shopping with us, have a nice day!")

else:
  searchResult = new_df[new_df['product_title'].str.contains((searchTerm),case=False)]
  searchdupl = searchResult.drop_duplicates(subset='product_title')
  searchResultSorted = searchdupl.sort_values(by=['recommend_score','product_title'],ascending=False).reset_index()
  searchResultSorted.index = np.arange(1, len(searchResultSorted) + 1)
  final_result = searchResultSorted.loc[:,['product_id', 'product_title','recommend_score']].head(10)

  #Replace average VADER score column as recommendation score
  final_result = final_result.rename(columns={'product_id': 'Product ID','product_title': 'Product Title',
                                              'recommend_score': 'Recommendation Score'})
  if len(final_result.index) < 1:
    print ("\nSorry! No results is found, please search again")
  else:
    display(final_result.head(10))
```

```
Finding an outdoors product?

Search for your outdoors product here (or enter 'exit' to exit):camping
```

	Product ID	Product Title	Recommendation Score
1	B00TQAWYFE	Fox Outfitters MicroDry Towel - Ultra Compact ...	9.571296
2	B008VGQLPS	Clark NX-250 Four-Season Camping Hammock	9.559211
3	B00LLH515A	Solo Stove Campfire - 4+ Person Compact Wood B...	9.488433
4	B00V2K00MU	[Durable Hammock & Strap Bundle] Serac Classic...	9.429545
5	B00PT110NG	Neolite Double Camping Hammock - Lightweight P...	9.405603
6	B00DQ3QS7C	Solo Stove 3 Pot Set - Stainless Steel Camping...	9.366071
7	B00KIGWNCE	Microfiber Travel Towel - Large 52" x 32" with...	9.344637
8	B00UVF4TZQ	SIERRA LEDs - Super Bright LED Camping Lantern...	9.335849
9	B00V97YCLQ	Eltronica LED Collapsible Camping Lantern	9.299074

Figure 4.15: Implementation of Search Engine. The search engine allows the user to search for their intended buying items and the results will be ranked by the recommendation score.

4.6 Recommendation System

The type of recommendation system used in this study can be considered as collaborative filtering, where explicit feedback (the product reviews) is extracted to generate recommendations based on the relevance of the product which is matched by the input search term and the words in the product title.

To further expand on the functionality of the system, another collaborative filtering is added on top of the search engine implemented. To extract the implicit feedback from the users, a customer-product matrix is used to generate recommendations to predict the customers' product preferences based on the other similar customers. The similarity between users is calculated by using a similarity matrix called cosine similarity, where the ratings are normalized by subtracting the mean.

4.6.1 Mapping customer ids and product ids into a matrix

```
from scipy.sparse import csr_matrix

def create_matrix(df):

    N = len(new_df['customer_id'].unique())
    M = len(new_df['product_id'].unique())

    # Map Ids to indices
    user_mapper = dict(zip(np.unique(new_df["customer_id"]), list(range(N))))
    product_mapper = dict(zip(np.unique(new_df["product_id"]), list(range(M))))

    # Map indices to IDs
    user_inv_mapper = dict(zip(list(range(N)), np.unique(new_df["customer_id"])))
    product_inv_mapper = dict(zip(list(range(M)), np.unique(new_df["product_id"])))

    user_index = [user_mapper[i] for i in new_df['customer_id']]
    product_index = [product_mapper[i] for i in new_df['product_id']]

    X = csr_matrix((new_df["vader_sentiment_score"], (product_index, user_index)), shape=(M, N))

    return X, user_mapper, product_mapper, user_inv_mapper, product_inv_mapper

X, user_mapper, product_mapper, user_inv_mapper, product_inv_mapper = create_matrix(new_df)
```

Figure 2.16: Mapping customer ids and product ids into a matrix

To conduct collaborative filtering, all the customer IDs and product IDs are mapped into a sparse matrix as shown in Figure 4.16. This matrix is also known as the user-item interaction matrix where it shows the interaction between the user (customer) to the item (product). This matrix represents the collaborative filtering contribution of the model.

4.6.2 Finding nearest neighbours using KNN

K-nearest neighbours (KNN) algorithm is a non-parametric supervised learning model where it can make inferences for new samples based on existing data points separated into different clusters. KNN can calculate the feature similarity distance between a target item with others in the database, hence returning K-nearest products as the most similar product recommendations. Figure 4.17 shows the implementation of KNN model building to find the k-nearest products with the target product.

```
from sklearn.neighbors import NearestNeighbors
"""
Find similar products using KNN
"""
def find_similar_products(product_id, X, k, metric='cosine', show_distance=False):

    neighbour_ids = []

    product_ind = product_mapper[product_id]
    product_vec = X[product_ind]
    k+=1
    kNN = NearestNeighbors(n_neighbors=k, algorithm="brute", metric=metric)
    kNN.fit(X)
    product_vec = product_vec.reshape(1,-1)
    neighbour = kNN.kneighbors(product_vec, return_distance=show_distance)
    for i in range(0,k):
        n = neighbour.item(i)
        neighbour_ids.append(product_inv_mapper[n])
    neighbour_ids.pop(0)
    return neighbour_ids


product_titles = dict(zip(new_df['product_id'], new_df['product_title']))
```

Figure 4.17: Building the KNN model

4.6.3 Combination of Collaborative Filtering

The advantage of this implicit collaborative filtering method with the KNN model is that it can capture intrinsic subtle attributes and the embeddings are learnt automatically. However, this technique has a cold start problem, which cannot handle fresh items. To deal with this problem, this recommendation system is only applied after the search engine, where the top 1 listing from the search results will be used as the subject in this recommendation system to generate other relevant products for the customer. By doing so, the implicit (KNN model) and explicit (sentiment score and star rating) are all included in the recommendation mechanism with the search engine (relevance) to help users search for the best products.

```
product_id = 'B004GD765O'

similar_ids = find_similar_products(product_id, X, k=10)
product_title = product_titles[product_id]

print(f"Since you found {product_title},")
print("\nYou may also like these:\n")
for i in similar_ids:
    print(product_titles[i])
```

```
Since you found Kelty Trail Ridge 4 Tent,

You may also like these:

Texsport Wayford 12' x 9' Portable Mesh Screenhouse Arbor Canopy for Backyard and Camping
RavX Drinker Can and Cup and Bottle Holder
Origin8 Pro Uno-S Saddle
Wenzel Camp-Away Airbed with Comfort Adjust Pump
MSR WhisperLite
Yaktrax Pro Traction Cleats for Walking, Jogging, or Hiking on Snow and Ice
Packtowl Nano Light Towel
Mountain House Breakfast Skillet
Selle Royal Respiro Moderate Saddle
Coleman Biscayne Big and Tall Warm Weather Sleeping Bag
```

Figure 4.18: Collaborative Filtering Recommendation System. By having the product id, the algorithm can generate recommendations through the KNN model and product-user matrix.

4.7 Summary

This chapter outlined the processes involved at every stage of the implementation of this study. The dataset used and the pre-processing process was discussed in detail. Less relevant variables and missing data were dropped from the dataset before running exploratory data analysis. Through exploratory data analysis, more insights were discovered from the data distribution charts of the dataset. The steps taken to perform sentiment analysis were described where the VADER library is used for sentiment scoring. After sentiment scoring, a search engine is created to help users find their products according to relevance by using a search engine to match their product title and the search term. Furthermore, more relevant items are recommended through an implicit feedback collaborative model. In summary, this chapter highlighted the techniques and adopted approaches at every stage of the study. The results from this implementation will be presented in the following chapter.

CHAPTER 5

RESULTS AND ANALYSIS

5.1 Introduction

This chapter presents the findings and results of the sentiment analysis of the product reviews in this project. Word clouds were generated to find the most frequently mentioned words in the reviews. The correlation analysis between the average sentiment score and the average 5-star rating was also done to find the relationship between both variables. The implementation of the search engine and recommendation system was also tested to ensure they are suggesting relevant items to the users. Lastly, a web application is created to present the search engine online to make it accessible through a web browser.

5.2 Word Cloud

To identify the factors that affect customers' online shopping experience, words that appear the most in the reviews are identified. This is done by tokenising the review text into individual word tokens and combining them into a word cloud. Word clouds represent a visual presentation of words by giving prominence to words that appear more frequently. Therefore, the words that were mentioned most by the customers is an indication of what customer cares about while considering purchase decisions.

Figure 5.1: Word Cloud. The words that are most frequently mentioned in the product reviews are visually represented in the word cloud.

The word ‘works great’ and ‘great product’ is part of the largest representations in the word cloud, which means that customers are mainly happy with the functions of the product that they purchased. Other than that, ‘well made’ also appear as one of the highest appeared word in the review. Hence, besides the functionality of the product, customers love the quality of the products, which is further validated with words like ‘good quality’, ‘good product’, ‘great quality’ and ‘high quality’ appearing in the word cloud. Surprisingly, the word ‘water bottle’ also featured a lot in the reviews by the customer. This finding is significant because the term is a noun instead of an adjective. Therefore, it has appeared that water bottles are a highly sought-after product in the outdoors category. On the other hand, compared to functionality and quality, the price aspect of the products is lesser mentioned in the reviews with only ‘great price’ featured notably inside the word cloud.

The concept of the iron triangle in the project management field (Pollack et al., 2018) presents that, between good quality, fast and cheap, only two aspects can be fulfilled for all cases. With regards to the product reviews, there is only a small representation of ‘fast shipping’ in the word cloud, which is in line with the iron triangle concept where fast shipping is seldom mentioned by customers after saying the quality and price are good for them.

Figure 5.2: Iron Triangle - Good, Fast, Cheap. The concept of the iron triangle in the project management field (Pollack et al., 2018), presents that between good quality, fast and cheap, only two aspects can be fulfilled for all cases.

5.3 Correlation between the average sentiment score and the average rating score

Besides finding the factors of purchase decisions, it is important to investigate the relationship between the text reviews and the actual star ratings given by the customers. Therefore, a correlation test is done to find the relationship between the sentiment scores and the star ratings. To perform this analysis, the mean of the sentiment scores and star ratings are calculated in the groups of product ids. Pearson's correlation coefficient is then used to calculate the correlation between the two variables.

Pearson's correlation coefficient is widely used as a measure to indicate the quality of bivariate connections between different variables. Hair et al. outlined the general principles of the coefficient extent and the quality of the connections (Hair et al, 2007, cited in Zamani et al., 2020).

Table 5.1: The scale of Pearson's Correlation Coefficient (Zamani et al., 2020)

Scale of correlation coefficient	Value
$0 < r \leq 0.19$	Very Low Correlation
$0.2 \leq r \leq 0.39$	Low Correlation
$0.4 \leq r \leq 0.59$	Moderate Correlation
$0.6 \leq r \leq 0.79$	High Correlation
$0.8 \leq r \leq 1.0$	Very High Correlation

Figure 5.3 shows the Pearson's correlation between the average rating and the average sentiment scores for the products. The correlation between the 2 variables is 0.667 with a significant level (p-value) of zero, which means that the relationship is correlated and both variables are relatively linked to each other. As a measure of the strength of the correlation, referring to Table 4.1, 0.667 lies between the range of 0.6 and 0.79, hence there is a high correlation between both variables. This indicates that generally, the sentiment of the customers who left reviews on the products is in line with the star rating scores that they left on the products.

```
# calculate Pearson's correlation
corr, _ = pearsonr(new_df['avg_rating'], new_df['avg_sentiment_score'])
print('Pearsons correlation: %.3f' % corr)

# calculate ttest significance
stats.ttest_ind(new_df['avg_rating'], new_df['avg_sentiment_score'])
```

```
Pearsons correlation: 0.667
Ttest_indResult(statistic=869.6214804286179, pvalue=0.0)
```

Figure 5.3: Correlation between the average rating and the average sentiment score. The correlation between the 2 variables is 0.667 with a significant level (p-value) of zero, which means that the relationship is correlated and both variables are relatively linked to each other.

However, on the other hand, this study only applies the sentiment scoring by using the VADER library. There are still many other methods to perform sentiment scoring like positive and negative word count normalization or libraries like Textblob and SentiWordNet. More comparisons between different methods can be done in future works.

5.4 Search Engine

As the final outcome, to help customers find their products, the search engine was combined with the collaborative filtering model to generate results for the search for their intended purchases. In the example shown in Figure 5.4, assuming one user is looking to buy an outdoors camping tent, the user used the search engine to search the keyword ‘tent’. The results listed rank the product ‘Pacific Breeze Easy Up Beach Tent’ as number 1 for the search term because it has the highest recommendation score compared to the others.

```
print('Finding an outdoors product?\n')
searchTerm=input("Search for your outdoors product here (or enter 'exit' to exit):")

if searchTerm.lower() == "exit":
  print("\n-----------------------------------------------\nThanks for shopping with us, have a nice day!")

else:
  searchResult = new_df[new_df['product_title'].str.contains((searchTerm),case=False)]
  searchdupl = searchResult.drop_duplicates(subset='product_title')
  searchResultSorted = searchdupl.sort_values(by=['recommend_score','product_title'],ascending=False).reset_index()
  searchResultSorted.index = np.arange(1, len(searchResultSorted) + 1)
  final_result = searchResultSorted.loc[:,['product_id', 'product_title','recommend_score']].head(10)

  #Replace average VADER score column as recommendation score
  final_result = final_result.rename(columns={'product_id': 'Product ID','product_title': 'Product Title',
                                              'recommend_score': 'Recommendation Score'})
  if len(final_result.index) < 1:
    print ("\nSorry! No results is found, please search again")
  else:
    display(final_result.head(10))
    print(final_result.iloc[0,0])

    product_id = final_result.iloc[0,0]
    similar_ids = find_similar_products(product_id, X, k=10)
    product_title = product_titles[product_id]

    print(f"Since you found the no.1 product: {product_title},")
    print("\nYou may also like these:\n")
    for i in similar_ids:
      print(product_titles[i])
```

```
Finding an outdoors product?

Search for your outdoors product here (or enter 'exit' to exit):tent
```

	Product ID	Product Title	Recommendation Score
1	B00RQQRTSM	Pacific Breeze Easy Up Beach Tent	9.391122
2	B004GD765O	Kelty Trail Ridge 4 Tent	9.345175

Figure 5.4: Searching for the term 'tent'

If the search term that the customer was finding cannot be found matching with the words inside the product title, no results will be shown, and the customer is instructed to search again with other keywords. This example can be seen in Figure 5.5 where the term ‘chelsea’ is searched. Other than that, when the word ‘exit’ is entered in the search term, the search engine takes it as an instruction to quit the searching process and a thank you message is displayed to the user. An example of exiting the searching process is displayed in Figure 5.6.

```
print('Finding an outdoors product?\n')
searchTerm=input("Search for your outdoors product here (or enter 'exit' to exit):")

if searchTerm.lower() == "exit":
  print("\n-----------------------------------------------\nThanks for shopping with us, have a nice day!")

else:
  searchResult = new_df[new_df['product_title'].str.contains((searchTerm),case=False)]
  searchdupl = searchResult.drop_duplicates(subset='product_title')
  searchResultSorted = searchdupl.sort_values(by=['recommend_score','product_title'],ascending=False).reset_index()
  searchResultSorted.index = np.arange(1, len(searchResultSorted) + 1)
  final_result = searchResultSorted.loc[:,['product_id', 'product_title','recommend_score']].head(10)

  #Replace average VADER score column as recommendation score
  final_result = final_result.rename(columns={'product_id': 'Product ID','product_title': 'Product Title',
                                              'recommend_score': 'Recommendation Score'})
  if len(final_result.index) < 1:
    print ("\nSorry! No results is found, please search again")
  else:
    display(final_result.head(10))
    print(final_result.iloc[0,0])

    product_id = final_result.iloc[0,0]
    similar_ids = find_similar_products(product_id, X, k=10)
    product_title = product_titles[product_id]

    print(f"Since you found the no.1 product: {product_title},")
    print("\nYou may also like these:\n")
    for i in similar_ids:
      print(product_titles[i])
```

```
Finding an outdoors product?

Search for your outdoors product here (or enter 'exit' to exit):chelsea

Sorry! No results is found, please search again
```

Figure 5.5: Searching for the term 'chelsea'

```
print('Finding an outdoors product?\n')
searchTerm=input("Search for your outdoors product here (or enter 'exit' to exit):")

if searchTerm.lower() == "exit":
  print("\n-----------------------------------------------\nThanks for shopping with us, have a nice day!")

else:
  searchResult = new_df[new_df['product_title'].str.contains((searchTerm),case=False)]
  searchdupl = searchResult.drop_duplicates(subset='product_title')
  searchResultSorted = searchdupl.sort_values(by=['recommend_score','product_title'],ascending=False).reset_index()
  searchResultSorted.index = np.arange(1, len(searchResultSorted) + 1)
  final_result = searchResultSorted.loc[:,['product_id', 'product_title','recommend_score']].head(10)

  #Replace average VADER score column as recommendation score
  final_result = final_result.rename(columns={'product_id': 'Product ID','product_title': 'Product Title',
                                              'recommend_score': 'Recommendation Score'})
  if len(final_result.index) < 1:
    print ("\nSorry! No results is found, please search again")
  else:
    display(final_result.head(10))
    print(final_result.iloc[0,0])

    product_id = final_result.iloc[0,0]
    similar_ids = find_similar_products(product_id, X, k=10)
    product_title = product_titles[product_id]

    print(f"Since you found the no.1 product: {product_title},")
    print("\nYou may also like these:\n")
    for i in similar_ids:
      print(product_titles[i])
```

```
Finding an outdoors product?

Search for your outdoors product here (or enter 'exit' to exit):exit

-----------------------------------------------
Thanks for shopping with us, have a nice day!
```

Figure 5.6: Exiting the search engine

5.5 Web App

Web application (web app) refers to an application program that is run on a remote server and can be accessed through the web browser interface. Streamlit is an open-source Python framework used to create web apps for data science and machine learning purposes. It is easy to deploy web apps using Streamlit because the apps can be written in a similar way to a python code.

To help the users easily access the search engine, a web app for the search engine was created and deployed by using Streamlit as shown in Figure 5.7 (Wei Aun, 2022). Through this web app search engine, users can easily search for relevant items and the results will be ranked by their recommendation scores. Besides, below the list of the search results, there are additional recommendations through collaborative filtering. Therefore, this web app essentially unifies the explicit (product reviews) and implicit (product-user interaction) feedback in the recommendation model by quantifying the sentiment of product reviews and generating additional recommendations from the results list.

E-Commerce Recommendations

Outdoors Category

Finding an outdoors product?

tent

Search Now

	Product ID	Product Title	Average Rating	Recommendation Score
1	B004GD765O	Kelty Trail Ridge 4 Tent	4.7895	0.8223
2	B000K7LU0M	Eureka! Copper Canyon 1512 - Tent (sleeps 12)	4.4156	0.8084
3	B000MAOEB4	ALPS Mountaineering Zephyr 2 Backpacking Tent	4.5385	0.8026
4	B002Q3LICS	Columbia Cougar Flats II Family Cabin Dome Tent	4.6604	0.8015
5	B004EQEB3I	ALPS Mountaineering Meramac 3 Tent	4.7458	0.7978
6	B00RQQRTSM	Pacific Breeze Easy Up Beach Tent	4.9102	0.7924
7	B000EQCWO2	Eureka! Sunrise 9 -Tent (sleeps 4-5)	4.5571	0.7920
8	B0007IS62E	Columbia Cougar Flats Six to Eight-Person Two-Room Cabin Tent	4.4590	0.7886
9	B00170JZE4	Columbia Cougar Flats II 15-Foot by 10-Foot 8 Person 2 Room Family Cabin Dome Tent	4.4815	0.7880
10	B000EJNGX0	Wenger Lugano 16- by 10-Foot Two-Room Eight-Person Family Tent with Canopy	4.6462	0.7840

Since you found the no.1 product: Kelty Trail Ridge 4 Tent,

You may also like these:

Texsport Wayford 12' x 9' Portable Mesh Screenhouse Arbor Canopy for Backyard and Camping

RavX Drinker Can and Cup and Bottle Holder

Origin8 Pro Uno-S Saddle

Wenzel Camp-Away Airbed with Comfort Adjust Pump

MSR WhisperLite

Yaktrax Pro Traction Cleats for Walking, Jogging, or Hiking on Snow and Ice

Packtowl Nano Light Towel

Mountain House Breakfast Skillet

Selle Royal Respiro Moderate Saddle

Coleman Biscayne Big and Tall Warm Weather Sleeping Bag

Figure 5.7: Web App interface (Wei Aun, 2022)

5.6 Summary

This chapter discussed the outcome of the sentiment analysis of the product reviews. The most frequent words in the reviews were discovered by using word clouds. A correlation analysis between the average sentiment score and the star rating was performed and the findings from the analysis were discussed in detail. After that, a search engine is created to help users find their relevant products by finding keywords through the product titles. Finally, the search engine was deployed into a web application on Streamlit, and the results were presented in this chapter.

CHAPTER 6

CONCLUSION AND RECOMMENDATIONS

6.1 Introduction

This chapter summarizes the whole progress of the project. Every stage of this project, like the data pre-processing, sentiment scoring, the search engine and the recommendations are briefly discussed, and conclusions are formed based on the outcome of this project. The findings from this project and its contribution are discussed further. To improve the contribution of the project, future recommendations are also suggested in this chapter.

6.2 Conclusions

In conclusion, this project was done by creating product recommendations based on sentiment scoring on the product reviews in an Amazon outdoors product dataset. Observations with missing data points were dropped and only verified purchases were used in further stages of the project. Next, sentiment scoring is done by using the natural language techniques and VADER library. After that, a search engine and recommendation system are created to help the customers find relevant products based on the product titles. Lastly, a web application is also deployed so that the search engine can be easily accessed through a web browser.

6.3 Importance and Contributions of the Study

This project outlined a framework to implement sentiment analysis from the customer reviews to the recommendation mechanism in e-commerce platforms. The framework can extract sentiment information from the customer reviews to help judge the quality of the products and hence act as an alternative scoring mechanism to rank products. It is also found that through this alternative scoring method, there is a strong correlation with the existing 5-star rating system, which indicates that this proposed method can supplement the existing system. This is important to help the e-commerce community to grow healthier with the help of text reviews instead of only 5-star ratings, with the help of this hybrid recommendation score implemented in this study. As the community grows, e-commerce companies can also reduce the need to incentivise customers to leave their reviews for the products they purchased.

6.4 Future Recommendations

As the recommendation method done in this project is only collaborative filtering, the effect of sales velocity and content-based filtering is yet to be explored. Other than that, the collaborative filtering recommendations were not measured for their accuracy because this project mainly focuses on laying out the framework of implementing sentiment analysis into the recommendation structure in the e-commerce industry. Furthermore, future works can include an optimization study of the weightages between the 5-star rating score and the sentiment scores. Additionally, the dates for the reviews can also be weighted because old reviews may not reflect the latest user opinions about a product.

REFERENCES

Alqaryouti, O., Siyam, N., Monem, A. A., & Shaalan, K. (2020). Aspect-based sentiment analysis using smart government review data. *Applied Computing and Informatics*. https://doi.org/10.1016/J.ACI.2019.11.003

Amalia, R., Bijaksana, M. A., & Darmantoro, D. (2018). Negation handling in sentiment classification using rule-based adapted from Indonesian language syntactic for Indonesian text in Twitter. *Journal of Physics: Conference Series*, *971*(1). https://doi.org/10.1088/1742-6596/971/1/012039

Amazon. (2016a). *What is Amazon Vine?* Amazon Website. https://www.amazon.com/vine/about

Amazon. (2016b, October 3). *Update on customer reviews*. The Amazon Blog. https://www.aboutamazon.com/news/innovation-at-amazon/update-on-customer-reviews

Asani, E., Vahdat-Nejad, H., & Sadri, J. (2021). Restaurant recommender system based on sentiment analysis. *Machine Learning with Applications*, *6*(June), 100114. https://doi.org/10.1016/j.mlwa.2021.100114

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, 107134. https://doi.org/10.1016/j.knosys.2021.107134

Cai, X. H., Liu, P. Y., Wang, Z. H., & Zhu, Z. F. (2017). Fine-grained sentiment analysis based on sentiment disambiguation. *Proceedings - 2016 8th International Conference on Information Technology in Medicine and Education, ITME 2016*, 557–561. https://doi.org/10.1109/ITME.2016.0132

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*(Mlm), 4171–4186.

Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data*, *2*(1). https://doi.org/10.1186/s40537-015-0015-2

Felfernig, A., Polat-Erdeniz, S., Uran, C., Reiterer, S., Atas, · Muesluem, Ngoc, T., Tran, T., Azzoni, P., Kiraly, · Csaba, & Dolui, · Koustabh. (2019). An overview of recommender systems in the internet of things. *Journal of Intelligent Information Systems*, *52*, 285–309. https://doi.org/10.1007/s10844-018-0530-7

Giannoulis, C. (2019). *Rescaling Sets of Variables to Be on the Same Scale. The Analysis Factor*. https://www.theanalysisfactor.com/rescaling-variables-to-be-same/

Hutto, C.J. and Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for. *Eighth International AAAI Conference on Weblogs and Social Media*, 18. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewPaper/8109

Infiniti Research. (2021). *Global E-Commerce Market 2021-2025*. ReportLinker. https://www.reportlinker.com/p04188481/Global-E-Commerce-Market.html?utm_source=GNW

Jebb, A. T., Parrigon, S., & Woo, S. E. (2017). Exploratory data analysis as a foundation of inductive research. *Human Resource Management Review*, *27*(2), 265–276. https://doi.org/10.1016/j.hrmr.2016.08.003

Karnan, T., & Seenuvasan, G. (2017). *Sentiment Analysis in E-Commerce*. *6*(8), 55–62.

Kauffmann, E., Peral, J., Gil, D., Ferrández, A., Sellers, R., & Mora, H. (2019). A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making. *Industrial Marketing Management*, *90*(November 2018), 523–537. https://doi.org/10.1016/j.indmarman.2019.08.003

Komorowski, M., Marshall, D. C., Salciccioli, J. D., & Crutain, Y. (2016). Exploratory Data Analysis. In *Secondary Analysis of Electronic Health Records* (pp. 185–203). Springer International Publishing. https://doi.org/10.1007/978-3-319-43742-2_15

Kontsewaya, Y., Antonov, E., & Artamonov, A. (2021). Evaluating the Effectiveness of Machine Learning Methods for Spam Detection. *Procedia Computer Science*, *190*(2019), 479–486. https://doi.org/10.1016/j.procs.2021.06.056

Kupiyalova, A., Satybaldiyeva, R., & Aiaskarov, S. (2020). Semantic search using natural language processing. *Proceedings - 2020 IEEE 22nd Conference on Business Informatics, CBI 2020*, *2*, 96–100. https://doi.org/10.1109/CBI49978.2020.10065

Laura Hautala. (2021). *Amazon's never-ending fake reviews problem, explained - CNET*. CNET. https://www.cnet.com/features/amazons-never-ending-fake-reviews-problem-explained/

Ligthart, A., Catal, C., & Tekinerdogan, B. (2021). Analyzing the effectiveness of semi-supervised learning approaches for opinion spam classification. *Applied Soft Computing*, *101*, 107023. https://doi.org/10.1016/J.ASOC.2020.107023

Liu, N. N., Xiang, E. W., Zhao, M., & Yang, Q. (2010). Unifying explicit and implicit feedback for collaborative filtering. *Proceedings of the 19th ACM International Conference on*

Information and Knowledge Management - CIKM '10, 1445. https://doi.org/10.1145/1871437.1871643

Mai, L., & Le, B. (2021). Joint sentence and aspect-level sentiment analysis of product comments. *Annals of Operations Research*, *300*. https://doi.org/10.1007/s10479-020-03534-7

Mukherjee, P., Badr, Y., Doppalapudi, S., Srinivasan, S. M., Sangwan, R. S., & Sharma, R. (2021). Effect of Negation in Sentences on Sentiment Analysis and Polarity Detection. *Procedia Computer Science*, *185*(June), 370–379. https://doi.org/10.1016/j.procs.2021.05.038

Niemi, J. (2015). *Amazon Reviews*. 1–24. https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt

Ofer, D., Brandes, N., & Linial, M. (2021). The language of proteins: NLP, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, *19*, 1750–1758. https://doi.org/10.1016/j.csbj.2021.03.022

Peng, Q., & Zhong, M. (2014). Detecting Spam Review through Sentiment Analysis. *Journal of Software*, *9*(8). https://doi.org/10.4304/jsw.9.8.2065-2072

Pollack, J., Helm, J., & Adler, D. (2018). What is the Iron Triangle, and how has it changed? *International Journal of Managing Projects in Business*, *11*(2), 527–547. https://doi.org/10.1108/IJMPB-09-2017-0107

Poria, S., Cambria, E., Winterstein, G., & Huang, G. Bin. (2014). Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems*, *69*(1), 45–63. https://doi.org/10.1016/j.knosys.2014.05.005

Ramachandran, D., & Parvathi, R. (2019). Analysis of Twitter Specific Preprocessing Technique for Tweets. *Procedia Computer Science*, *165*, 245–251. https://doi.org/10.1016/j.procs.2020.01.083

Salgado, C. M., Azevedo, C., Proença, H., & Vieira, S. M. (2016). Missing Data. In *Secondary Analysis of Electronic Health Records* (pp. 143–162). Springer International Publishing. https://doi.org/10.1007/978-3-319-43742-2_13

Sasikala, P., & Mary Immaculate Sheela, L. (2020). Sentiment analysis of online product reviews using DLMNN and future prediction of online product using IANFIS. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00308-7

Shen, R. P., Zhang, H. R., Yu, H., & Min, F. (2019). Sentiment based matrix factorization with reliability for recommendation. *Expert Systems with Applications*, *135*, 249–258. https://doi.org/10.1016/j.eswa.2019.06.001

Shopee Malaysia. (2021a). *[Product Rating] How do I rate the product I purchased?* https://help.shopee.com.my/my/s/article/How-do-I-rate-the-product-I-purchased

Shopee Malaysia. (2021b). *[Product Rating] Why is my product review removed?* https://help.shopee.com.my/s/article/why-is-my-product-review-removed

Silva, R. M., Almeida, T. A., Cardoso, E. F., & Silva, R. M. (2018). Towards automatic filtering of fake reviews. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2018.04.074

Sivapalan, S., Sadeghian, A., Rahnama, H., & Madni, A. M. (2014). Recommender systems in e-commerce. *World Automation Congress Proceedings*, *August*, 179–184. https://doi.org/10.1109/WAC.2014.6935763

Tang, D. (2015). Sentiment-specific representation learning for document-level sentiment analysis. *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 447–451. https://doi.org/10.1145/2684822.2697035

Wei Aun, O. (2022). *Streamlit*. https://share.streamlit.io/weiaun96/ecommerce-recommendation-web-app/main/web_app.py

Zamani, N., Bahrom, N. A., Meor Fadzir, N. S., Mohd Ali @ Mohd Fauzy, N. S., Anuar, N. F., Rosman, S. A., Sivam, S., Muthutamilselvan, K., & Isai, K. I. A. (2020). A Study on Customer Satisfaction Towards Ambiance, Service and Food Quality in Kentucky Fried Chicken (KFC), Petaling Jaya. *Malaysian Journal of Social Sciences and Humanities (MJSSH)*, *5*(4), 84–96. https://doi.org/10.47405/mjssh.v5i4.390

APPENDIX A

ETHICAL APPROVAL OF RESEARCH PROJECT

Office Record	Receipt – APU Fast-Track Ethical Approval
Date Received:	Student name: Student number: Received by:
Received by:	Date:

APU/APIIT FAST-TRACK ETHICAL APPROVAL FORM (STUDENTS)

Tick one box (level of study):
- ☑ POSTGRADUATE (PhD / MPhil / Masters)
- ☐ UNDERGRADUATE (Bachelors degree)
- ☐ FOUNDATION / DIPLOMA / Other categories

Tick one box (purpose of approval):
- ☐ Thesis / Dissertation / FYP project
- ☐ Module assignment
- ☑ Other: Capstone Project

Title of Programme on which enrolled … … Data Science and Business Analytics

Tick one box: ☑ Full-Time Study or ☐ Part-Time Study

Title of project / assignment … Application of Sentiment Analysis in E-Commerce Recommendation System

Name of student researcher … … Ong Wei Aun

Name of supervisor / lecturer… … Mr. Raheem Mafas

Student Researchers- please note that certain professional organisations have ethical guidelines that you may need to consult when completing this form.

Supervisors/Module Lecturers - please seek guidance from the Chair of the School Research Ethics Committee if you are uncertain about any ethical issue arising from this application.

		YES	NO	N/A
1	Will you describe the main procedures to participants in advance, so that they are informed about what to expect?			✓
2	Will you tell participants that their participation is voluntary?			✓
3	Will you obtain written consent for participation?			✓
4	If the research is observational, will you ask participants for their consent to being observed?			✓
5	Will you tell participants that they may withdraw from the research at any time and for any reason?			✓
6	With questionnaires and interviews will you give participants the option of omitting questions they do not want to answer?			✓
7	Will you tell participants that their data will be treated with full confidentiality and that, if published, it will not be identifiable as theirs?			✓
8	Will you give participants the opportunity to be debriefed i.e. to find out more about the study and its results?			✓

If you have ticked No to any of Q1-8, you should complete the full Ethics Approval Form.

		YES	NO	N/A
9	Will your project/assignment deliberately mislead participants in any way?			✓
10	Is there any realistic risk of any participants experiencing either physical or psychological distress or discomfort?			✓
11	Is the nature of the research such that contentious or sensitive issues might be involved? This includes research which could induce psychological stress, anxiety or humiliation, or cause more than minimal pain.			✓
12	Does your research involve the use of sensitive materials? Eg, records of personal or sensitive confidential information,			✓

13	Does your research require external agency approval?			✓
14	Does your research use hazardous or controlled substance?			✓
15	Does your research require you to visit participants in their home or non-public space?			✓
16	Does your research use genetically modified organisms?			✓
17	Does your research investigate illegal activities or behaviours?			✓
18	Does your research involve discussion or collection of information on potentially sensitive, embarrassing or distressing topics, administrative or secure data? This includes research involving respondents through internet where visual images are used, and where sensitive issues are discussed			✓
19	Does your research involve invasive or potentially intrusive procedures?			✓
20	Does your research involve administration of substances?			✓
21	Will your research be involved in the collection/ processing of human tissue samples			✓
22	Will your participants be receiving financial compensation for participating in your research?			✓
23	Will your research data be used in the future after the conclusion of your project?			✓
24	Will your research involve in processing sensitive data belonging to an organisation/persons?			✓
25	Will your research be collecting photographs, videos, and audio recordings of the participants?			✓
26	Will the participants' personal particulars be known to any third party?			✓
27	Will the participants' data confidentiality be made known to the public?			✓
28	Will the research be conducted where the safety of the researchers maybe in question?			✓
29	Will be the research be conducted outside of the UK and/or Malaysia?			✓
30	Will your research involve human participants at premises other than those of the University?			✓

If you have ticked Yes to any of Q9 – 30, you should complete the full Ethics Approval Form. In relation to question 10 this should include details of what you will tell participants to do if they should experience any problems (e.g. who they can contact for help). You may also need to consider risk assessment issues.

			YES	NO	N/A
31	Does your project/assignment involve work with animals?				✓
32	Do participants fall into any of the following special groups? **Note that you may also need to obtain satisfactory clearance from the**	Children (under 18 years of age)			✓
		People with communication or learning difficulties			
		Patients			
		People in custody			
		People who could be regarded as vulnerable or lack capacity to make decision for themselves			
		People engaged in illegal activities (eg drug taking)			
		Groups of people whose relationship among each other allow one to have influence over the other such as: Carers and patients with chronic conditions; teachers and their students; prison authorities and prisoners;			

	relevant authorities	employers and employees			
		Deceased person's body parts or other human tissues including bodily fluids (e.g. blood, saliva).			
		groups where permission of a gatekeeper is normally required for initial access to members.			
		Human participants who are off-campus APU staff or students who wish to carry out investigations involving human participants at premises other than those of the University			
33	Does the project/assignment involve external funding or external collaboration where the funding body or external collaborative partner requires the University to provide evidence that the project/assignment had been subject to ethical scrutiny?				✓

If you have ticked Yes to any Q31-33**,** you should complete the full Ethics Approval Form. There is an obligation on student and supervisor to bring to the attention of the APU School Research Ethics Committee any issues with ethical implications not clearly covered by the above checklist.

STUDENT RESEARCHER
Provide in the boxes below (plus any other appended details) information required in support of your application. THEN SIGN THE FORM.

Please Tick Boxes

I consider that this project/assignment has no significant ethical implications requiring a full ethics submission to the APU School Research Ethics Committee.	✓
I am aware of APU liability policy and will make the necessary arrangement for insurance coverage of all researchers and participants of the project/assignment.	✓
Give a brief description of participants, procedure of recruitment and procedure of data collection (methods, tests used etc) in up to 150 words. The data used in the project is made available by Amazon, where the licence rights allow a limited, non-exclusive, non-transferable, non-sublicensable, revocable license to access and use the Reviews Library for purposes of academic research.	
I also confirm that: i) All key documents e.g. consent form, information sheet, questionnaire/interview, and all material such as emails and posters for the purpose of recruitment of participants are appended to this application.	✓
Or ii) Any key documents e.g. consent form, information sheet, questionnaire/interview schedules which need to be finalised following initial investigations will be submitted for approval by the project/assignment supervisor/module lecturer before they are used in primary data collection.	✓

Signed… … … … … … … … … … … … … Print Name… … … … … … … … … … … … … Date… … … …
(Student Researcher)

Ong Wei Aun

19 Mar 2022

Within this document, any variation to the items considered which affects ethical issues of the stated research will require submission of a revised research plan and research methodology details; as a consequence, new ethical consent may need to be sought.
The completed form (and any attachments) should be submitted for consideration by your Supervisor/Module Lecturer

SUPERVISOR/MODULE LECTURER
PLEASE CONFIRM THE FOLLOWING:

	Please Tick Box
I consider that this project/assignment has no significant ethical implications requiring a full ethics submission to the APU School Research Ethics Committee	✓
I have checked and approved the key documents required for this proposal (e.g. consent form, information sheet, questionnaire, interview schedule)	✓(N/A)

SUPERVISOR AND SECOND ACADEMIC SIGNATORY

STATEMENT OF ETHICAL APPROVAL (please delete as appropriate)

1) **THIS PROJECT/ASSIGNMENT HAS BEEN CONSIDERED USING AGREED APU PROCEDURES AND IS NOW APPROVED**

2) **THIS PROJECT/ASSIGNMENT HAS BEEN APPROVED IN PRINCIPLE AS INVOLVING NO SIGNIFICANT ETHICAL IMPLICATIONS, BUT FINAL APPROVAL FOR DATA COLLECTION IS SUBJECT TO THE SUBMISSION OF KEY DOCUMENTS FOR APPROVAL BY SUPERVISOR (see Appendix A)**

Signed… RMF … Print Name… Mr. Raheem Mafas … Date… 20.04.2022 …
(Supervisor/Lecturer)

Signed… LGR … Print Name… Prof. Dr. R.Logeswaran … Date… 21/04/2022 …
(Second Academic Signatory)

Office Record Date Received: Received by:	Receipt – Appendix A (APU Fast-Track Ethics Form) Student name: Student number: Received by: Date:

APPENDIX A
AUTHORISATION FOR USE OF KEY DOCUMENTS

Completion of Appendix A is required when for good reasons key documents are not available when a fast track application is approved by the supervisor/module lecturer and second academic signatory.

I have now checked and approved all the key documents associated with this proposal e.g. consent form, information sheet, questionnaire, interview schedule

Title of project/assignment… …

… …......................

Name of student researcher …..…….

Student ID: ……………………………… Intake: … … … … … … … … … … … … … … … ……..

Signed… … … … … … … … … … … … Print Name… … … … … … … … … … … … Date… … ……... ……..
(Supervisor/Lecturer)

APPENDIX B

LOG SHEETS FOR SUPERVISORY SESSION

(APU: Serial Number)

PLS V1.0

Project Log Sheet – Supervisory Session

Notes on use of the project log sheet:

1. This log sheet is designed for meetings of more than 15 minutes duration, of which there must be at minimum SIX (6) during the course of the project (SIX mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant sections of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. The student should leave a copy (after the session) of the Project Log Sheet with the supervisor and to the administrator at the academic counter. A copy is retained by the student to be filed in the project file.
6. It is recommended that students bring along log sheets of previous meetings together with the project file during each supervisory session.
7. The log sheet is an important deliverable for the project and an important record of a student's organisation and learning experience. The student **must** hand in the log sheets as an appendix of the final year documentation, with sheets dated and numbered consecutively.

Student's name: ONG WEI AUN **Date:** 15 Oct 2021 **Meeting No:** 1

Project title: APPLICATION OF SENTIMENT ANALYSIS IN E-COMMERCE RECOMMENDATION SYSTEMS

Supervisor's name: MR. RAHEEM MAFAS **Supervisor's signature:** RMF

Items for discussion (noted by student before mandatory supervisory meeting):

1) Review on RMCP report
2) How to prepare for CP1

Record of discussion (noted by student during mandatory supervisory meeting):

- References better be maximum of 4 years old
- follow the Research Proposal Template / guidance provided by the Module Lec

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

- Find references for the CP1
- Start doing literature review

Note: A student should make an appointment to meet his or her supervisor (via the consultation system) at least ONE (1) week prior to a mandatory supervisor session – please see document on project timelines. In the event a supervisor could not be booked for consultation, the project manager should be informed ONE (1) week prior to the session so that a meeting can be subsequently arranged.

(APU: Serial Number)

PLS V1.0

Project Log Sheet – Supervisory Session

Notes on use of the project log sheet:

1. This log sheet is designed for meetings of more than 15 minutes duration, of which there must be at minimum SIX (6) during the course of the project (SIX mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant sections of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. The student should leave a copy (after the session) of the Project Log Sheet with the supervisor and to the administrator at the academic counter. A copy is retained by the student to be filed in the project file.
6. It is recommended that students bring along log sheets of previous meetings together with the project file during each supervisory session.
7. The log sheet is an important deliverable for the project and an important record of a student's organisation and learning experience. The student **must** hand in the log sheets as an appendix of the final year documentation, with sheets dated and numbered consecutively.

Student's name: ONG WEI AUN **Date:** 10 Nov 2021 **Meeting No:** 2

Project title: APPLICATION OF SENTIMENT ANALYSIS IN E-COMMERCE RECOMMENDATION SYSTEMS

Supervisor's name: MR. RAHEEM MAFAS **Supervisor's signature:** RMF

Items for discussion (noted by student before mandatory supervisory meeting):

1) Potential topic for CP1

Record of discussion (noted by student during mandatory supervisory meeting):

- Video on Instagram stating the problem with online shopping and product reviews
- CP1 scope: extracting sentiment from product reviews and creating recommendations for user

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

- Start focusing on the past literature about sentiment analysis of product reviews.

Note: A student should make an appointment to meet his or her supervisor (via the consultation system) at least ONE (1) week prior to a mandatory supervisor session – please see document on project timelines. In the event a supervisor could not be booked for consultation, the project manager should be informed ONE (1) week prior to the session so that a meeting can be subsequently arranged.

(APU: Serial Number)

PLS V1.0

Project Log Sheet – Supervisory Session

Notes on use of the project log sheet:

1. This log sheet is designed for meetings of more than 15 minutes duration, of which there must be at minimum SIX (6) during the course of the project (SIX mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant sections of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. The student should leave a copy (after the session) of the Project Log Sheet with the supervisor and to the administrator at the academic counter. A copy is retained by the student to be filed in the project file.
6. It is recommended that students bring along log sheets of previous meetings together with the project file during each supervisory session.
7. The log sheet is an important deliverable for the project and an important record of a student's organisation and learning experience. The student **must** hand in the log sheets as an appendix of the final year documentation, with sheets dated and numbered consecutively.

Student's name: ONG WEI AUN **Date:** 1 Dec 2021 **Meeting No:** 3

Project title: APPLICATION OF SENTIMENT ANALYSIS IN E-COMMERCE RECOMMENDATION SYSTEMS

Supervisor's name: MR. RAHEEM MAFAS **Supervisor's signature:** RMF

Items for discussion (noted by student before mandatory supervisory meeting):

1) Cannot find LR/official sources stating about this problem about irrelevant product reviews
2) 10k word requirement for CP1

Record of discussion (noted by student during mandatory supervisory meeting):

- Can use business decision makers' opinion or statements supporting that the problem statement is existing and to be solved even though cannot find any published journal articles because some problems may be new and yet to be explored.
- 10k words is not a must

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

- Find resources from websites like Amazon or Shopee
- Preparing CP1 report and presentation

Note: A student should make an appointment to meet his or her supervisor (via the consultation system) at least ONE (1) week prior to a mandatory supervisor session – please see document on project timelines. In the event a supervisor could not be booked for consultation, the project manager should be informed ONE (1) week prior to the session so that a meeting can be subsequently arranged.

(APU: Serial Number)

PLS V1.0

Project Log Sheet – Supervisory Session

Notes on use of the project log sheet:

1. This log sheet is designed for meetings of more than 15 minutes duration, of which there must be at minimum SIX (6) during the course of the project (SIX mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant sections of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. The student should leave a copy (after the session) of the Project Log Sheet with the supervisor and to the administrator at the academic counter. A copy is retained by the student to be filed in the project file.
6. It is recommended that students bring along log sheets of previous meetings together with the project file during each supervisory session.
7. The log sheet is an important deliverable for the project and an important record of a student's organisation and learning experience. The student **must** hand in the log sheets as an appendix of the final year documentation, with sheets dated and numbered consecutively.

Student's name: ONG WEI AUN **Date:** 7 Mar 2022 **Meeting No:** 4

Project title: APPLICATION OF SENTIMENT ANALYSIS IN E-COMMERCE RECOMMENDATION SYSTEMS

Supervisor's name: MR. RAHEEM MAFAS **Supervisor's signature:** RMF

Items for discussion (noted by student before mandatory supervisory meeting):

1) CP1 report feedback, show edited update for CP2 report – what is Ethics Form?
2) Research plan (gantt chart) was missing for CP1 - Is it needed to be included for CP2? In chp 3 or appendix?
3) EDA, Created sentiment scoring using VADER and positive/negative words
   - how to count weighted rating for rankings? Can just make a sum of them with original star rating?
4) Progress Update: sentiment scoring almost done, next is recommendation system

Record of discussion (noted by student during mandatory supervisory meeting):

- Apspace Masters - send email to program leader for ethics form
- Research plan – last part of chapter 1 introduction
- Just take sentiment scoring alone

Action List (to be attempted or completed by student by the <u>next</u> mandatory supervisory meeting):

Note: A student should make an appointment to meet his or her supervisor (via the consultation system) at least ONE (1) week prior to a mandatory supervisor session – please see document on project timelines. In the event a supervisor could not be booked for consultation, the project manager should be informed ONE (1) week prior to the session so that a meeting can be subsequently arranged.

Project Log Sheet

(APU: Serial Number)

PLS V1.0

Project Log Sheet – Supervisory Session

Notes on use of the project log sheet:

1. This log sheet is designed for meetings of more than 15 minutes duration, of which there must be at minimum SIX (6) during the course of the project (SIX mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant sections of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. The student should leave a copy (after the session) of the Project Log Sheet with the supervisor and to the administrator at the academic counter. A copy is retained by the student to be filed in the project file.
6. It is recommended that students bring along log sheets of previous meetings together with the project file during each supervisory session.
7. The log sheet is an important deliverable for the project and an important record of a student's organisation and learning experience. The student **must** hand in the log sheets as an appendix of the final year documentation, with sheets dated and numbered consecutively.

Student's name: ONG WEI AUN **Date:** 28 Mar 2022 **Meeting No:** 5

Project title: APPLICATION OF SENTIMENT ANALYSIS IN E-COMMERCE RECOMMENDATION SYSTEMS

Supervisor's name: MR. RAHEEM MAFAS **Supervisor's signature:** RMF

Items for discussion (noted by student before mandatory supervisory meeting):

1) Is there any problem with the fast-track ethics form? Haven't got signature
2) Since I created a search results-based recommendation, how can we measure the performance of the recommendation system? Or can just show some demo of using it?
3) Created a web app for this implementation, should include inside chapter 5 results also? https://share.streamlit.io/weiaun96/ecommerce-recommendation-web-app/main/web_app.py

Record of discussion (noted by student during mandatory supervisory meeting):

- Revise the fast-track form, change Yes to N/A
- Check sentiment score convert into recommendation score
- Build recommendation models based on sentiment scoring

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

- Collaborative filtering is based on feedback from customers, can be implicit and explicit. Sentiment score derived from reviews are actually explicit feedback from customers.

Note: A student should make an appointment to meet his or her supervisor (via the consultation system) at least ONE (1) week prior to a mandatory supervisor session – please see document on project timelines. In the event a supervisor could not be booked for consultation, the project manager should be informed ONE (1) week prior to the session so that a meeting can be subsequently arranged.

(APU: Serial Number)

PLS V1.0

Project Log Sheet – Supervisory Session

Notes on use of the project log sheet:

1. This log sheet is designed for meetings of more than 15 minutes duration, of which there must be at minimum SIX (6) during the course of the project (SIX mandatory supervisory sessions).
2. The student should prepare for the supervisory sessions by deciding which question(s) he or she needs to ask the supervisor and what progress has been made (if any) since the last session, and noting these in the relevant sections of the form, effectively forming an agenda for the session.
3. A log sheet is to be brought by the STUDENT to each supervisory session.
4. The actions by the student (and, perhaps the supervisor), which should be carried out before the next session should be noted briefly in the relevant section of the form.
5. The student should leave a copy (after the session) of the Project Log Sheet with the supervisor and to the administrator at the academic counter. A copy is retained by the student to be filed in the project file.
6. It is recommended that students bring along log sheets of previous meetings together with the project file during each supervisory session.
7. The log sheet is an important deliverable for the project and an important record of a student's organisation and learning experience. The student **must** hand in the log sheets as an appendix of the final year documentation, with sheets dated and numbered consecutively.

Student's name: ONG WEI AUN **Date:** 18 Apr 2022 **Meeting No:** 6

Project title: APPLICATION OF SENTIMENT ANALYSIS IN E-COMMERCE RECOMMENDATION SYSTEMS

Supervisor's name: MR. RAHEEM MAFAS **Supervisor's signature:** RMF

Items for discussion (noted by student before mandatory supervisory meeting):

1) CP1 part showed high Turnitin, need to edit?
2) Can send draft report to check before actual submission?
3) Presentation date, duration 10mins?

Record of discussion (noted by student during mandatory supervisory meeting):

- No need worries about CP1 Turnitin
- Check with Dr. Logeswaran, because supervisor will be away for 3 weeks

Action List (to be attempted or completed by student by the next mandatory supervisory meeting):

Note: A student should make an appointment to meet his or her supervisor (via the consultation system) at least ONE (1) week prior to a mandatory supervisor session – please see document on project timelines. In the event a supervisor could not be booked for consultation, the project manager should be informed ONE (1) week prior to the session so that a meeting can be subsequently arranged.