# 

**KABARAK UNIVERSITY**

**SCHOOL OF COMPUTER SCIENCE, ENGINEERING AND TECHNOLOGY**

**DEPARTMENT OF COMPUTER SCIENCE AND IT**

**PROJECT PROPOSAL**

**INTE 424 IT PROJECT I**

**PROJECT TITLE:**

SENTIMENT ANALYSIS ON PRODUCT REVIEW FOR GENUINE RATING AND RECOMMENDATION OF PRODUCTS BASED ON COLLABORATIVE FILITERING.

**A Project Report Documentation Submitted in The Department of Computer Science and IT in partial fulfillment of the degree of Information Technology**

**PRESENTED BY:**

MUNYOKI DENIS MWANZIA - INTE/MG/1947/05/18

**SUPERVISOR:**

MR. JOSHUA MUTAI

**SUBMITTED ON:**

2ND DECEMBER 2021

# DECLARATION

The work contained in this project proposal has not been previously presented in any other higher education institutions. To the best of my knowledge, ability and belief the document contains no material previously published or written by another person except where due references are made.

Student’s Name: DENIS MWANZIA

Student’s Signature: ………………………………. [Signature]

...………………………......... [Date]

# APPROVAL

I confirm that the work report in this project proposal was carried out by student under my supervision.

It was submitted to Kabarak University Department of Computer Science and IT.

Supervisor’s Name: MR. JOSHUA MUTAI

Supervisor’s Signature: ………………………………. [Signature]

...………………………......... [Date]

# DEDICATION

I dedicate this project to God who has been my strength and divine inspiration in everything I do.

I would also like to dedicate this project to my parents Joel Munyoki and Ndanu, my brother and my sisters for their support and encouragement.

I would also like to dedicate this project to my classmates for their support and help within the period of my academic journey at Kabarak University

# ACKNOWLEDGEMENT

I would like to express my sincerest gratitude to my supervisor Mr. Mutai for his professional guidance, motivation, immense knowledge, variable advice and continuous support that has led to successful completion of this project proposal. I am so grateful for his reliability that has enabled me to consult with him to enhance the quality of my project proposal.

I also acknowledge the support of my fellow classmates who have either directly or indirectly contributed to success of this project proposal, I am so grateful for your constant help and encouragement throughout the project.

# ABSTRACT

Sentiment analysis is a rapidly emerging domain in the area of research in the field of Natural Language Processing (NLP). It has gained much attention in recent years and especially with the Ecommerce companies. These ecommerce companies are using sentiment analysis to analyze or rather extract opinions from the reviews that their customers give about their products. It is no surprise that most of ecommerce companies have increased their rate of revenue generation in the last five years. Sentiment analysis of product reviews has helped the ecommerce companies to find out their customers preferences and interest to their products and therefore, it has made it easier to provide personalized systems where their customers can view more of the products of their interest. In this project I am going to develop a sentiment analysis model which will be able to determine whether one’s review is either a positive or negative review and the model will be integrated to an ecommerce site made up of Django (python web framework). The dataset used in the training of the model will be from a data repository know to be Kaggle. Words such as adjectives and adverbs will be able to convey opposite sentiment with the help of negative prefixes. Support Vector Machine (SVM) will be the machine learning algorithm that will be used in the training of sentiments analysis model. The performance of the model will be evaluated through evaluation measures such as accuracy and precision score. This sentiment analysis mode will be able to randomly assign rating to a review based on its sentiment polarity that is rating of 1-2 for negative reviews and rating of 4-5 for positive reviews. A recommender system will be developed and also integrated to the same ecommerce system where it will utilize the randomly assigned rating to recommend products to the customer. This ecommerce system framework will be utilized by the willing SMEs who have not yet shifted to ecommerce. This framework will give them same competitive advantages as the other organizations which have already adopted ecommerce. As a result, the citizens of Kenya will be able to get products of high quality and at customer friendly price.

# TABLE OF CONTENTS

[DECLARATION I](#_Toc89057325)

[APPROVAL I](#_Toc89057326)

[DEDICATION II](#_Toc89057327)

[ACKNOWLEDGEMENT III](#_Toc89057328)

[ABSTRACT IV](#_Toc89057329)

[TABLE OF CONTENTS V](#_Toc89057330)

[TABLE OF FIGURES VII](#_Toc89057331)

[ABBREVATIONS AND SYMBOLS VIII](#_Toc89057332)

[CHAPTER ONE 1](#_Toc89057333)

[1.0 INTRODUCTION 1](#_Toc89057334)

[1.2 BACKGROUND OF THE STUDY 3](#_Toc89057335)

[1.2.1 User rating Vs product Quality. 5](#_Toc89057336)

[1.2.2 The power of sentiment analysis and recommendation systems in electronic commerce. 6](#_Toc89057337)

[1.3 STATEMENT OF THE PROBLEM 7](#_Toc89057338)

[1.4 PURPOSE OF THE STUDY 8](#_Toc89057339)

[1.5 OBJECTIVES 8](#_Toc89057340)

[1.6 RESEARCH QUESTIONS 9](#_Toc89057341)

[1.7 SIGNIFICANCE OF THE STUDY 9](#_Toc89057342)

[1.8 EXPECTED OUTCOME 10](#_Toc89057343)

[1.9 SCOPE OF THE STUDY 10](#_Toc89057344)

[1.10 LIMITATIONS OF THE STUDY 11](#_Toc89057345)

[1.11 SCALABILITY 11](#_Toc89057346)

[CHAPTER TWO 12](#_Toc89057347)

[2.0 INTRODUCTION 12](#_Toc89057348)

[2.1 SYSTEM REVIEWS 13](#_Toc89057349)

[2.1.1 Sentiment Analysis and Opining Mining by G. Vinodhini and RM. Chandrasekaran (June 6,2012). 13](#_Toc89057350)

[2.1.2 Sentiment Categorization with Machine Learning Techniques by Andre’s Cassinelli, Chih-Wei Chen (June 5,2009). 13](#_Toc89057351)

[2.1.3 Twitter as a corpus for sentiment analysis and opinion mining by Pak, A., & Paroubek, P. (2010, May) 14](#_Toc89057352)

[2.1.4 Types of recommender systems 14](#_Toc89057353)

[2.2 INTERGRATED SUMMARY OF LITERATURE STUDIED 19](#_Toc89057354)

[CHAPTER THREE 20](#_Toc89057355)

[3.0 INTRODUCTION. 20](#_Toc89057356)

[3.1 SOFTWARE DEVELOPMENT LIFE CYCLE. 20](#_Toc89057357)

[3.1.1 Original methodologies. 22](#_Toc89057358)

[3.2.2 Agile methodologies. 24](#_Toc89057359)

[3.2.2.1 Scrum methodology. 25](#_Toc89057360)

[3.2.2.2 Extreme methodology. 25](#_Toc89057361)

[3.3 SELECTION OF METHODOLOGY. 25](#_Toc89057362)

[3.4 JUSTIFICATION OF SELECTED METHODOLOGY. 26](#_Toc89057363)

[3.5 LOCATION OF STUDY. 27](#_Toc89057364)

[3.6 RESEARCH DESIGN. 27](#_Toc89057365)

[3.7 TARGET POPULATION. 27](#_Toc89057366)

[3.8 SAMPLING TECHNIQUE AND SAMPLE SIZE. 27](#_Toc89057367)

[3.9 DATA COLLECTION TOOL. 28](#_Toc89057368)

[3.9.1 Questionnaire. 28](#_Toc89057369)

[3.9.2 Observation. 29](#_Toc89057370)

[3.10 SYSTEM DESIGN. 29](#_Toc89057371)

[3.10.1 Functional decomposition diagram. 29](#_Toc89057372)

[3.10.2 Use case diagram 30](#_Toc89057373)

[3.10.3 class diagram. 31](#_Toc89057374)

[3.10.4 Activity diagram: 32](#_Toc89057375)

[REFERENCES. 34](#_Toc89057376)

# TABLE OF FIGURES

[Figure 1. Top 5 e-commerce leading countries worldwide. 3](#_Toc89028999)

[Figure 2. Revenue generation of e-commerce in Kenya and projected revenue up to 2025. 4](#_Toc89029000)

[Figure 3 Agile Development (Branden 2018) 26](#_Toc89029001)

[Figure 4 Functional Decomposition Diagram 29](#_Toc89029002)

[Figure 5 Use Case Diagram 30](#_Toc89029003)

[Figure 6 Class Diagram 31](#_Toc89029004)

[Figure 7 Activity Diagram for Users 32](#_Toc89029005)

[Figure 8 Activity Diagram for Admin 33](#_Toc89029006)

# ABBREVATIONS AND SYMBOLS

|  |  |
| --- | --- |
| SME | Small and Medium Enterprises |
| NLP | Natural Language Processing |
| SA | Sentiment Analyzer |
| SVM | Support Vector Machine |
| SDLC | Software Development Life Cycle |
| TF-IDF | Term Frequency-Inverse Document Frequency |
| RM | Recommender System |
|  |  |
|  |  |
|  |  |
|  |  |

# CHAPTER ONE

## 1.0 INTRODUCTION

Sentiment analysis is an approach in NLP which is used to extracts opinions or emotions towards an item or an event. With increased number and popularity of e-commerce technology most companies and people are trying to sell and buy products online. This means that many people have shifted to online store shopping as compared to physical store shopping. This is as result of reliability in ecommerce platform. For instance, one can order a product of their desire irrespective of their location and time. In Kenya things are not different. According to UNCTAD B2B Commerce index, Kenya was ranked 88th in fastest growing e-commerce economy worldwide and 4th rank in Sub Saharan Africa (behind Mauritius, Nigeria and South Africa). This is as a result of mobile connectivity which has improve access to internet and online payments to facilitate transactions for example M-Pesa, T-Cash, Airtel Money and Equitel. With all the mentioned infrastructure in place and the increased number of internet users in Kenya; Statista estimates Kenya’s e-commerce market to reach $1.7 billion in 2021. In Kenya there only two dominant e-commerce platform that is Kilimall and Jumia. With predicted increase in online shopping; these two dominant e-commerce platforms will be able to provide the products in demand but not at consumer-friendly price. With many SMEs reluctant on shifting to e-commerce due to high price in owning such a platform, you will find that the dominant e-commerce platforms have higher competitive advantage compared to them (SMEs) hence the prices of their (Jumia and Kilimall) products are probably going to be high because there are only the two of them competing against each other. Some of key competitive advantages that is two dominant e-commerce systems have over the SMEs are knowing the preferences of their customers and being able to get reviews of what their customers think of their products and services. This is all possible with sentiment analysis of the product reviews which can categories a review as either a positive review or negative review and with that they are able to get rating which can be used in recommendation of products to the customer. In this project proposal I propose Sentiment analysis model which will be used to extract the opinions of customers through the analysis of reviews gives on certain product. These reviews will be categorized as either a positive or negative review. The model will randomly allocate a rating that is 1-2 for negative reviews and 4-5 for negative reviews. These ratings will be used by recommender system to recommend products to customers that making it a personalized e-commerce system and therefore meeting one of web 3 properties. This framework will be used by the interested SMEs who will be able to catch up with other enterprises such as Jumia and kilimall. This will also increase competition and as result the prices of products will be consumer friendly.

## 1.2 BACKGROUND OF THE STUDY

The internet has become a very integral part of modern-day society and its rapid adoption has given rise to concept and practice of e-commerce (electronic commerce), which has become a fundamental necessity for each organization in the word today (Alyoubi,2015). As a result, many people and individuals are trying to buy goods and services online. Most of these transactions are being conducted over the internet. Nowadays more and more people are shifting from physical store shopping model to an online model. This is because e-commerce has made it easy, fast and convenient and thus making it a compelling allure to customer who prefer the easy and the efficiency way of ordering and delivery of products and services. According to global survey carried Perset, K. (2010) found out that some products bought online are universal. Some of these leading products are books, electronic equipment, clothes, shoes, cosmetics etc. According to Statista (2021), the global e-commerce market is 4.28 trillion US dollars in 2020 and it is projected to grow to 5.4 trillion in 2022.

E-commerce is booming in the developed and advance countries of the world and this can be seen in the figure below.

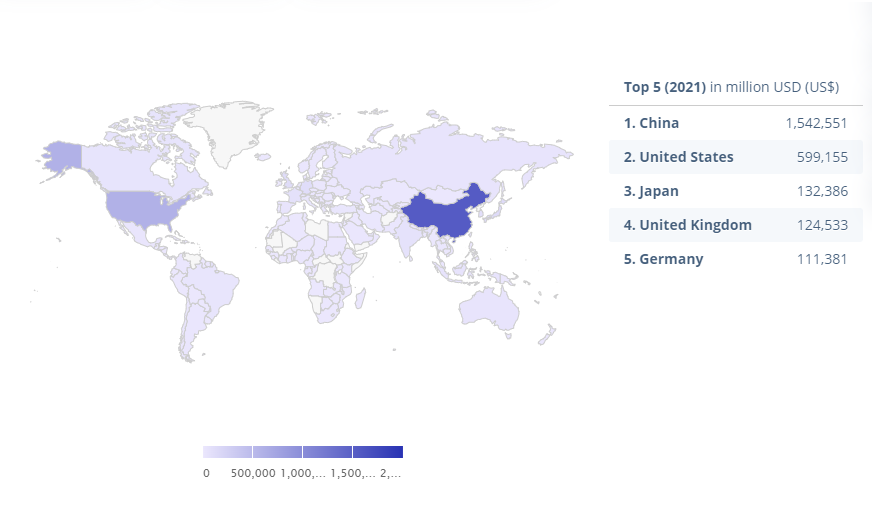


Figure 1. Top 5 e-commerce leading countries worldwide.

In Kenya e-commerce growth is picking up despite the fact we have few e-commerce systems selling products on demand. This can be seen in the figure below.

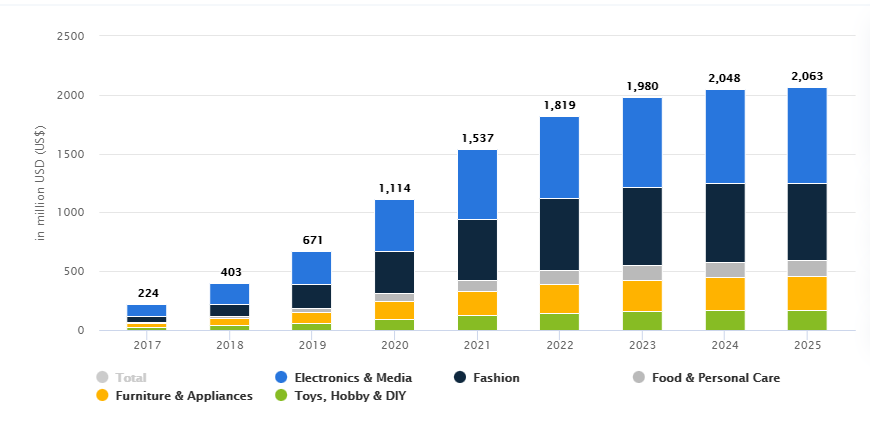


Figure 2. Revenue generation of e-commerce in Kenya and projected revenue up to 2025.

Unfortunately, despite e-commerce advanced technology and acceptance by large firms, the level of acceptance is not evidence amongst SMEs (Knol & Stroeken, 2010). For instance, in Kenya there are no many e-commerce platforms where the citizen can purchase the products of their desire. There are only two dominating e-commerce platform which include Jumia and Kilimall. For the time being these two dominant e-commerce platforms will be able to provide the product on demand but with increase demand in future they might be able but not in customer friendly prices. This will be as result of small and medium enterprises (SMEs) reluctant to shift e-commerce. Most of SMEs claim that it is too expensive to own and run e-commerce of their own. According to Barclay’s research, more than a fifth (20%) of brick and mortal retailers choose not to sell online due to concerns about cost of managing deliveries and returns.

Some of competitive advantages that these two dominating e-commerce systems in Kenya have over the SMEs include the ability to collect product or service reviews from their customers and the ability to recommend products to their customers. According to Matt Ellis 2020, no review or feedback has led to failure of successful e-commerce business but rather it results to improvement and success of the business.

It is one thing when you talk about how great your products are and it’s another whole thing when other shoppers talk about your products. According to Craig Bloem, 84% of people trust online reviews as much as personal recommendation from a friend and 64% make a purchase decision after reading one to six reviews. A lot of consumers when searching online for something to buy will take a look at an online review or rating for a product (Adrian R. Camilleri). It seems like a great way to get an over view on the quality of a product. As a result, average user rating of products has become a significant factor in driving sales across many product categories.

### 1.2.1 User rating Vs product Quality.

Customers use online rating because they make assumptions that they provide a good indicator of quality and service, For Example; a product with five-star or four-star rating is expected by the customer to be of higher quality unlike compared to a product with a three star. As per Journal Consumer Research (JCR) vol 45 issued on 6th April 2016 page no 817-833, a group of researchers tested this assumption where an impressive dataset comprising of 344,157 Amazon.com rating for 1272 products in 120 categories. The researchers found that the average user rating correlated poorly with the scores from consumer report, that is; a product with an average rating of four out five stars would only be objective 65% not 100% of the time compared to the other product with an average of three out of five stars.

### 1.2.2 The power of sentiment analysis and recommendation systems in electronic commerce.

Sentiment analysis or rather opinion mining is Natural language processing (NLP) approach which defines emotions or opinions and attitude towards a product or an item in a text format. This is an existing field which many e-commerce platforms have invested in. with help of sentiment analysis many e-commerce systems have been able to analyze the feedbacks from their customers about their product which are given inform of reviews. It is through the insights discovered through opinion mining that the e-commerce firm can improve on the products and services which received negative reviews. This results to satisfaction of their customers hence their growth and success.

Recommendation system can be defined as a model which is developed using machine learning algorithms and its purpose is to recommend items based on specified criteria. There are different types of recommendation systems which are classified according to the criteria which they are based on. For instance, we have recommendation systems based on item similarities and recommendation systems based on collaborative filtering. With a recommender system integrated to an e-commerce system is one of success driver for such e-commerce platform. For example, with a recommender system based on collaborative filtering can improve user experience (UX) as it will be able to provide a personalized recommendation of products to each user that uses the e-commerce system.

In conclusion both sentiment analysis of product reviews and recommender systems are the key driving power for growth and success of e-commerce systems. Bravo! to our ecommerce systems which have invested heavily on recommender system and opinion mining of product reviews but their still does not make it easy for customers when working with reviews that is they do not separate the positive reviews and negative or even indicate and therefor, the customers have to go through all reviews before making a decision from which store to order their products.

## 1.3 STATEMENT OF THE PROBLEM

In our country Kenya there a are limited number of e-commerce platform where the citizens can get their desired products. As results the existing e-commerce platform offer their products and services to the users but not at user friendly prices. Most of SMEs are reluctant to shift to e-commerce because it is expensive to own such platform. According to OCED (2006a) report, SMEs form the most dominant type of business, by which it accounts for over 95% and up to 99% of enterprises depending on country and job creation in OCED countries by SMEs is between 60-70%. These observations are in tandem with survey carried out by Government of Kenya (2010) which established that 96% of all entrepreneurs are SMEs operators, employ 95% of workforce and contributes to 18% to the Gross Domestic Product in Kenya.

There are considerable number of studies that have been carried to analyze various aspects of ecommerce in developing countries and developed countries including Kenya. Is turns out that there is an emerging gap in past research studies carried out. Most of these studies are focused on the factors influencing e-commerce adoption by SMEs (Mansour, & Al-Nawafleh, 2009; Shemi,2012; Wanjau et el., 2013; Ochola,2013) barriers or rather obstacles facing e-commerce systems (Kinuthia & Akinnusi, 2014).

As mentioned above its clear that most of research studies are focused on the adoption and barriers of e-commerce systems in Kenya, however, the solutions to the actual problems facing e-commerce systems seems to be lacking. For instance, the solution to the problem on the fact that SMEs are reluctant to adopt e-commerce because it is expensive is not addressed. This proposal project will address the problem on the high cost in owning an e-commerce system by answering the key questions, what are the factors leading to reluctance of SMEs to adopt e-commerce? How are existing e-commerce performing? and most importantly, what are the key drive powers for success of any ecommerce?

## 1.4 PURPOSE OF THE STUDY

The purpose of this study is to:

1. Study about B2C types of ecommerce systems in Kenya, sentiment analysis and recommendation system types that can actually be used in ecommerce systems.
2. Provide an online platform where users can view, buy, rate and review different products.
3. Provide a platform for SMEs to manage their products and sales online.
4. Build a recommender system that can recommend products to each and every individual user in the system.

## 1.5 OBJECTIVES

The objectives of this project are as follows:

1. To collect the requirements of an ecommerce system.
2. To design an ecommerce system.
3. To implement ecommerce system.
4. To test the ecommerce system.

## 1.6 RESEARCH QUESTIONS

1. How will the requirements of the system be gathered?
2. How will the system be designed?
3. How will the system be implemented?
4. How will the system be tested?

## 1.7 SIGNIFICANCE OF THE STUDY

The significance of this study is to provide information on the current status of e-commerce in Kenya, factors leading to SMEs reluctance to adopt e-commerce, a review different algorithm and techniques to extract feature wise summary of a product and analyze it to form an authentic review and build a powerful sentiment analysis and recommender models which will be integrated to an e-commerce framework/system for interested SMEs to start selling their products and services online. With the mentioned e-commerce frameworks, the operations for both seller and buyer will be made easier. For instance, the sellers will be able to view analyzed or rather categorized reviews from their customers and improve their services and products while the buy will experience an automated recommendation of products making their shopping experience great. With the proposed framework in place and operating the citizens of Kenya will be able to get product of their desire from multiple platforms and at customer friendly price.

## 1.8 EXPECTED OUTCOME

If and when the proposed system is adopted by SMEs the problem of high cost to own an e-commerce will be solve on the SMEs side and they will be able to enjoy the same competitive advantage like the existing e-commerce firms. With its existence the proposed system problem on customer friendly prices for the products of their desire will be solved due to competition and multiple platforms where customers can compare prices on different platform before making a purchase. This is system is also expected to help achieve the projected market of 1819 million US dollars by 2022.

## 1.9 SCOPE OF THE STUDY

As the main objective of this study is to develop and test an e-commerce system and to integrate it with a sentiment analysis and recommender system into the web-based application. This system’s functionality is divided into two main areas of concerns. The first area on scope of the proposed system is to analyze the review submitted by the customer, categorize it as either a positive while assigning it a random value between 0 and 3 or a negative review while assigning it a random value between 3 and 6 and not to validate whether the review is genuine or fake. These randomly assigned values will be used as a rating of specific product from each review. The other area on scope of the study is that the study is not specific to one region or country.

## 1.10 LIMITATIONS OF THE STUDY

The limitation to the study is that system may not work well with newly signed users due to the fact that the system is based on collaborative filtering where the recommendations emerge from a correlation between the intended user and other users, focused on collection of ratings. This therefore makes it impossible to judge the user and categorize a user with few ratings which is directly proportional not reviewing products which will in turn automatically rating of the product. This system only starts recommending products to the user after the user reviews some products.

## 1.11 SCALABILITY

The proposed system can be customized and adopted by every business model may it be an online book selling store, online pharmaceutical stores, online food stores, fashion (clothes and shoes) selling store etc.

# CHAPTER TWO

## 2.0 INTRODUCTION

This chapter present a wide range of literature review from the previous studies or rather research reports. It mainly focuses on the knowledge and ideas established in sentiment analysis and their evolution, their strengths and weaknesses as well as recommender systems and types of existing recommender systems.my research proposal topic is not a new topic but an active area of research. Both sentiment analysis (SA) and recommender system (RM) are popular fields of studies among the researchers and scientist since last few decades. Both SA and RM are fairly common in large tech companies such as Amazon and Google. With increased popularity of product reviews, people have been expressing their opinions or sentiment on the products. Customer’s opinion has a great impact on other customer decision making (Gretze & Yoo, 2008). Online product reviews may have another additional attribute such as rating (satisfaction score). These scores are usually given on score or range of one to five. For instance, a customer wants to buy a product he/she will go through the reviews given by others and arrive at conclusion. The probability of a customer buying the product increases with the number of positive reviews the product has received. In this era of information overload, it becomes difficult for users to get the information that they are really interested in and for content creators it becomes difficult for them to deliver their content to target population. This is the same case with e-commerce systems without a recommender system. With daily technological advancement in e-commerce systems, it is therefore important to put the SA and RM in the frontline since they become very important engines for branding and marketing. They both helps build trust, loyalty and typically describe what sets your product apart from the others. Hybrid system consisting of SA and RM can be used to address, deeper analysis of product reviews and solve the issues and problems of both buyers and sellers by automating the activity of product recommendation based on quality of user-item communication (Melville & Sindwani, 2018). Both SA and RM are reshaping or transforming e-commerce systems by providing the customers with both relevant information and suggestions which help customers to choose the relevant product. Hybrid system of SA and RM definitely increases sale figures as it offers customers with more varied products which improves user satisfaction and fully considers customer needs. It helps build trust, loyalty and typically describe what sets your services apart from the others. There are various tools that are available to analyze sentiments and recommend products. Some of existing online environments available for recommendation of products include content-based filtering, collaborative filtering, popularity based and hybrid approach (Anantha & Bhattula,2017). By going through research papers and report related to my research proposal topic, I have found very different techniques, classifications and approaches that are used different researchers and are key for success of my project.

## 2.1 SYSTEM REVIEWS

### 2.1.1 Sentiment Analysis and Opining Mining by G. Vinodhini and RM. Chandrasekaran (June 6,2012).

increasing in size. The sources of this data include social network, news, entertainment, reviews, blogs, discussions forums. Data analyzed focus on these opinions for sentiment analysis work. Researchers and data scientist are working on building a software to detect and classify the texts available online. The resulting information can give us a lot of information about our customers likes and disklike, what they want to buy and what they do not want to buy and what they do not want to buy. This information can be used by third party companies such as e-commerce to provide better services and deals to their customers or to help their customers to get better deals in terms of reviews. This paper also details on the survey describing about the techniques, approaches and methods in data analysis and the problem exist in the area of sentiment analysis.

### 2.1.2 Sentiment Categorization with Machine Learning Techniques by Andre’s Cassinelli, Chih-Wei Chen (June 5,2009).

This is a summary to calculate the sentiment of given review or text or opinion. It notes that methods of data analysis are nearly the same to the past works in data analysis in reviews. By applying classification technique on the data, it firstly starts by using a section data as training set to train itself and use the rest of data (test data) to evaluate itself. The approach mentioned in this paper describes the relationship between the objects in more efficient way.

### 2.1.3 Twitter as a corpus for sentiment analysis and opinion mining by Pak, A., & Paroubek, P. (2010, May)

Summarizes today’s social network sites like twitter, Facebook, LinkedIn etc. as famous tools to communicate among people on internet. Millions of people share information with each other. This information can be of use to some of people or a waste for some people. If this information is properly analyzed it could become a very important asset to any organization. This in information may be inform of opinions and could results to others. The author in this paper identifies these social networks sites very effective tools in generation of information about so many aspects in today’s life for humans. But there is litter done on analysis of the information these social network sites come to existence some years ago. The author specifies the details using twitter, one of most famous and used social network in present world; to explain the work of sentiment analysis.

### 2.1.4 Types of recommender systems

There are three approaches of recommendation systems; that is content based, collaborative filtering and hybrid filtering. Each of these approaches have their benefits and limitations as shown in later part of this section. Collaborative filtering is further divided into two main types: memory-based filtering and model-based filtering techniques. Memory-based is further divided into user-user based collaborative filtering and item-item based collaborative filtering. Model-based approach is further divided into clustering, Bayesian, association and neural network. On other hand hybrid filtering is combination of both content-based filtering and collaborative filtering.

#### 2.1.4.1 Content-based filtering

Content-based filtering, also referred as cognitive filtering, is used to recommend similar products that the customer had already bought from the system. It can also be used to recommend products to users based on other factors such as gender, age, geography, reviews, and usage patterns. According to Pazzani & Billsus (2015), preferences of any given customer can be determined by retrieval method such as term Frequency-Inverse Document Frequency (TF-IDF), cosine similarity matrix, long distance affair or by using machine learning techniques such as Naïve Bayes, Support Vector machine (SVM), Decision Trees etc.

##### 2.1.4.1.1 Cosine similarity matrix

This technique makes use of angle calculation of two vectors in order to estimate similarities between two items. Cosine similarity can be calculated as:

According to Sitikhu, et al. (2019), its decision is not based on magnitude but rather based on orientation.

##### 2.1.4.1.2 Term Frequency-Inverse Document Frequency

TF-IDF approach is used to process data and extract knowledge from the data. According to Beel, et al. (2017), Term frequency as the name suggest, involves measuring the frequency at which a word occur and can be calculated as ratio of total number of particular words occurs in a document and total number of words in the document. TF-IDF measure the significance of term in a document and can be calculated as:

IDF (an instance of word) = log\_e (total no. of documents / no pf documents with particular word in it)

Can be calculated as:

This approach can be used and effective only in the field where there is more textual data.

#### 2.1.4.2 collaborative filtering

The term collaborative filtering was first introduced by developers at xerox. Since its introduction it has developed and enhanced by using different technologies and algorithms. As the main aim of any recommender system is to recommend items to users that is based on their desires; collaborative filtering (CF) is used to suggest the item which are preferred by other related users. For collative filtering to suggest items there are a number of algorithms which have to be implement. Each algorithm has different set of features. According to HERLOCKER, et al. (2004), an algorithm can take more time as compared to another to better its functionality in particular dataset. Some of algorithms used are discussed below.

##### 2.1.4.2.1 Nearest neighborhood approach

K-nearest neighbor algorithm can be used to predict the user’s desire. It also keeps into consideration the existing users from the system. In this technique, groups of similar users are segregated and their averages ratings are calculated. In this case collaborative recommendation operate on assumption that similar customers like similar products. Therefore, the main task in this case is to find similar users from the set of all users in the database system. There are two ways that we can apply to find similar customers and that is; using ratings and reviews given by customers to different products and the other way is to track all the user’s activity in the web application. For recommendation to occur in this case we first need to find similarity between separated user and active user is calculated by using different techniques which are discussed below.

* + - * 1. **Pearson correlation**

This similarity approach was introduced in 1994 by Group Len. It is considered to be the best algorithm to find the similarity between the users. For instance, let the profile of the targeted user be x (in form of a vector) and the profile of another user be x (x ∈ T). In this case the score of similarity between the target user x and the neighbor y can be calculated by using the following formula.

Where, 𝑥𝑖 is the rating given by user to item i,

n is the total number of items and

𝑥̅ is the mean rating given by a user and calculated by using the following formula:

The value of 𝑅 𝑥, 𝑦 comes between -1 and +1. If a value is equal to -1, there is a negative correlation between two values and if the value is equal to or near to 1 there is positive correlation. Here x and y are the user whose similarity needs to be computed. (Katarya & verma, 2016)

* + - * 1. **Euclidean Distance**

We can use Euclidean distance to find the similarity between two customers. The distance between two customers can be calculated as:

Where xi and yi are the score of products given by two user u and v for same product n.

(Sondur & Chigadani, 2016)

* + - * 1. **Cosine Distance**

The similarities between user can be computed by taking ratings of users and putting two user’s rating as vector in an n-dimensional space and computing the cosine of angle between the two users. This can be shown by the equation below:

where w(a,u) is the measure of similarity between user a and the active user u.

i is the set of items rated by user a and u

Rui is the rating given by user u to item i and r is the mean rating given by user i

+1(same direction) and -1(opposite direction) represents high correlation and 0(irrelevant) represents no correlation. This approach gives both negative ranking and blank rated product as zero rating. (Liu, et al., 2014)

* + - * 1. **Adjusted Cosine Similarity**

This technique is able to recommend products to customers based upon the rating of products. The problem cosine distance is that customers product rating varies from one product to another. The solution to this problem is to subtract the average user rating from rating provided by user in each product. Therefore, the adjusted cosine similarity is calculated by using the formula below:

Where, u and v are the two customers. u is the customer to be predicted and v is the active customer in the system.

##### 2.1.4.2.2 Clustering based Collaborative filtering

This approach makes use of clusters of similar users. In simple terms clustering based collaborative filtering makes use of similar users based on their preferences. There are different types of clustering which include; distributed-based clustering, connectivity-based clustering, centroid-based clustering, density-based clustering etc. Any of mentioned distance above can be used for measurement of user similarity distance. The distance can be calculated by using the formula below:

Where 𝑃𝑎,𝑖 is the prediction of user a to item i. 𝐴(𝑟̅̅al) is the average rating of user a to product I and 𝐴(𝑟̅̅kl) is the average ratings of user k to item I. (Kim & Yang, 2015)

## 2.2 INTERGRATED SUMMARY OF LITERATURE STUDIED

Sentiment analysis is one of popular topics that researchers are currently working on. There are lot of research going on and especially on different languages that have not been studied until now such as Arabic, Hindi, Thai etc. There are various open-source libraries available for different programing languages like python and R which make it easy and possible to process and analyze text. Sentiment analysis can be used for various purposes for instance, reviewing product of a company, reviewing movies, expressing feeling or emotions of citizens towards their country etc. it is the most popular way to get information on social media and to analyze the collected information. To transform the collected information into something meaningful a classification technique must be used. The data must be in readable format, in English. Classifiers are used to tokenize and classify the data. The Super Wised learning technique is used with machine learning approach to detect sentiments and analyze the sentiments of the rest of the text. Supervised learning is linguistic approach in which text is first tokenized into tokens and added with tags to evaluate the sentiments of the text. There are various sources of data to evaluate which include: Social sites (Facebook.com, Twitter.com, LinkedIn.com), News websites and Movie Reviewing Sites and Products selling sites (Flipkart, Snapdeal, Amazon Blogs etc.) Techniques used presently are: Machine learning (Naïve Bias classifier, Support Vector Machine).

# CHAPTER THREE

## 3.0 INTRODUCTION.

A research methodology can be defined as the study of methods by which we gain knowledge. It usually comprises the cognitive processes imposed on research by the problems arising from the nature of its subject matter. This chapter introduces the methodologies and techniques that will play a very importance role in data collection and analysis of the collected data to aid in achieving the objectives of the proposed system. This section will also describe the software development life cycle, the existing methodologies, justification of selected methodology, research design, location of study, research target population, the sample size and data collection tools.

## 3.1 SOFTWARE DEVELOPMENT LIFE CYCLE.

Software development life cycle (SDLC) can be defined as framework that describes a systematic approach to software development. It aims at producing high quality software which meets the needs of end user. It includes some phases as discussed below.

**Phase I: Requirement Analysis and specifications**

Requirements are the needs of the customers/user that must be met to solve their problem and to achieve the objectives of developing the system. Requirement analysis involves studying the requirements of the customer or user to arrive at definition of requirements. Specification refers to description of essential requirements of the system and external interface. This task is performed system analyst. The system analyst holds a meeting with the stakeholders from the problem domain and tries to bring out as much information as possible. The requirement can be collected conducting interviews with the people in area of problem domain, studying the existing systems or even using questionnaires. The end product of this phase is Software Requirement Specifications (SRS). SRS served as reference document and it can be used to describe the requirement of the users of the proposed system.

**Phase II: Design**

SRS is the input for this phase. This phase starts with SRS which contains the requirement of the proposed system. The requirements are then mapped into an architecture. The architecture defines the components, interfaces and behavior of the system. The end product or rather the deliverable design document of this phase is the architecture.

**Phase III: Implementation**

System design in the architecture document from the design phase is the input of this phase. This phase is also referred as coding phase which will involve development of actual system by using high level programing language like python and other web technologies such as HTML, CSS and JavaScript. Each system design from the previous phase is implement independently through coding and later integrated to make the whole system. The deliverable of this phase is the product itself.

**Phase IV: Testing**

After the system is successfully implemented through coding, testing is conducted to evaluate whether it meets its requirements. Both black box testing and white box testing can used to test the developed system. Black box testing includes stress testing, acceptance testing and beta testing while white box testing involve unit testing, integration testing, system testing and regression testing.

**Phase V: Deployment.**

This means deploying the system to the actual environment where it will be used. For instance, this can be done by installing the software product to the system where the product will be used or even hosting the application on online such as Heroku where it can be accessed and used.

**Phase VI: Maintenance.**

The documentation of the software is required for its maintenance. Software maintenance can be into three. These categories include corrective maintenance for correcting the errors in the software, adaptive maintenance for the software to adopt to changing environment and perfective maintenance for making sure that the software meets the changing needs of its users.

### 3.1.1 Original methodologies.

#### 3.1.1.1 Waterfall model.

It is also referred as sequential model. It works on the principle ‘work to completion’ that is output of one phase becomes input in the next phase. All phases one after the other in linear sequence.

**When waterfall model is most appropriate**

1. When product definition is stable.
2. When the requirements are well documented, clear and fix.
3. When the project is short.
4. When there are no ambiguous requirements.

**Advantages of waterfall model**

1. Simple and easy to understand and use.
2. Well understood milestone.
3. Effective control.
4. Easy to manage due to rigidity of the model.

#### 3.1.1.2 Iterative model.

This SDLC model leads software development process in iterations. This model delivers full system in the beginning. The software product is developed in small scale and all the steps to be taken into considerations are followed. Each iteration result to software product which is complete in itself and has more features and functionalities than the previous one.

**When applicable.**

1. When the requirements of the system are clearly defined and understood.
2. When new technology is being used and being learnt by the development team while working on the project.
3. When there is high risk features and goals which may change in the future.

**Advantages of iterative model.**

1. Results are obtained early and periodically.
2. Progress can be measured.
3. Quality of the software product can be measured.
4. It supports changing requirements.

**Disadvantages of iterative model.**

1. More resources may be required.
2. High management complexity.
3. Not suitable when working with small projects.

#### 3.1.1.3 Spiral model.

Combines features of both iterative model and one of SDLC model and then adds risk analysis. It starts with determining the objectives and constraints of the software product at the start of iteration one.

**When to apply spiral model.**

1. When the customer is not sure of their requirement.
2. When requirements are complex and they need evaluation to get clarity.
3. When there is budget constraints and risk analysis is important.

**Advantages of spiral model.**

1. Frequent feedback from customers.
2. Risk can be dealt with less cost.
3. Changing requirements can be accommodated.

**Disadvantages of spiral model.**

1. Complex to implement and understand.
2. Limited to advanced skills.
3. Time consuming.
4. Scope creep.

## 3.2.2 Agile methodologies.

This is type of incremental model in which each subsequent release is built on and it adds functionality to previous release. It breaks system into small components which are implemented and delivered in incremental and rapid cycle. Each increment is tested to ensure software quality is maintained. Agile is needed when there are new changes that are needed to be implemented at low cost. Agile methodology has the following advantages.

1. A working product is delivered quickly and frequently.
2. Has technical excellence and good design and even late changes in requirements are welcome

### 3.2.2.1 Scrum methodology.

With this methodology the scrum teams work in iterations (sprints) that takes two weeks to at most one month. Changes cannot be made into these sprints. Once the sprint planning meeting is completed and a commitment to deliver a set of products is made, the set of products remain unchanged through the end of the sprint. Scrum product owner prioritizes the backlog of the product but the teams determine the sequence of the backlog items.

### 3.2.2.2 Extreme methodology.

Extreme is an agile framework that aims at producing software products of high quality. The customer states their requirements of the software to be developed. The software product is developed in small fragments and are frequently tested throughout the software development process. Customer acceptance test is performed at each stage to ensure the right product is developed. The source code of the product only includes code that is necessary to come up with the right product.

## 3.3 SELECTION OF METHODOLOGY.

The methodology that will be used in the development of the proposed system is agile methodology. Specifically extreme programming (XP)will be used. XP, unlike the other methodologies is very opinionated when it comes to engineering practice and it is built upon values and principles. Agile methodology development has aspects of prototyping. Prototyping is a worthwhile requirement collection technique involve building a prototype and then refine the other requirement to make final product.



Figure Agile Development (Branden 2018)

## 3.4 JUSTIFICATION OF SELECTED METHODOLOGY.

The reason for selection of agile is because of the following reasons.

1. A rapid delivery of working software product which brings about satisfaction to users.
2. New requirements are welcomed.
3. Minimum resources are required.
4. Regular adoption to changing requirements.
5. Working software is delivered frequently.

## 3.5 LOCATION OF STUDY.

The study will be conducted in Eastleigh market, Uhuru market and Gikomba all located in Nairobi Kenya. The rationale for selecting these areas is due to availability of large population to collect the require information, intensive observations therefore easy to access information and at minimum cost.

## 3.6 RESEARCH DESIGN.

This research problem will employ both the qualitative and quantitative form. The reason for choice of these two research designs is because they vary in strength and weakness and therefor the study will use both of them for complementally purpose. This study will use qualitative design to large extent quantitative design to lesser extent because the research exploratory in nature that is the research will use open ended and probing questions where the respondent will be required to give response in their own words. This method will be used to capture frequency of respondents and average level of response.

## 3.7 TARGET POPULATION.

According to Mugenda & Mugenda (2003), target population can be defined as as the population set which the researcher studies, and whose findings are used to generalize the general population. The target population of this study are the SMEs at Eastleigh, Uhuru market and Gikomba market in Nairobi County who have not yet adopted ecommerce system.

## 3.8 SAMPLING TECHNIQUE AND SAMPLE SIZE.

According to Webster (1985), a sample can be defined as a finite part of target population whose characteristics are studied for the purpose of getting information about the whole target population. To determine a sample size, we need to employ a specific technique. The sample size for this study will be determined by sample size determination formula advanced by Krejcie and Morgan (1970). The formula is given shown below.

Where n= Sample size

X2 = Chi-square for the specified confidence level at 1 degree of freedom

N= population size

P = population proportion

ME = Desired Margin of Error (expressed as a proportion)

## 3.9 DATA COLLECTION TOOL.

The requirements for proposed system will be collected using the following tools.

### 3.9.1 Questionnaire.

The questionnaire will contain both open ended and closed ended questions. The questionnaire will be constructed based on the research objectives. Open-ended questions which elicit qualitative data on subjective thoughts and different responses related to why they (SMEs) are reluctant to adopt ecommerce as well as characteristic of expected proposed system. Questionnaires as a data collection method will provides a relatively cheap and more efficient way of obtaining large amounts of information from a larger population.

## 3.9.2 Observation.

Observation as the term sound, it refers to collection of data through observations. Observation will be used to examine numerous existing web-based ecommerce applications similar to proposed ecommerce system. This will help to make the proposed ecommerce system as much practical as real world.

## 3.10 SYSTEM DESIGN.

### 3.10.1 Functional decomposition diagram.

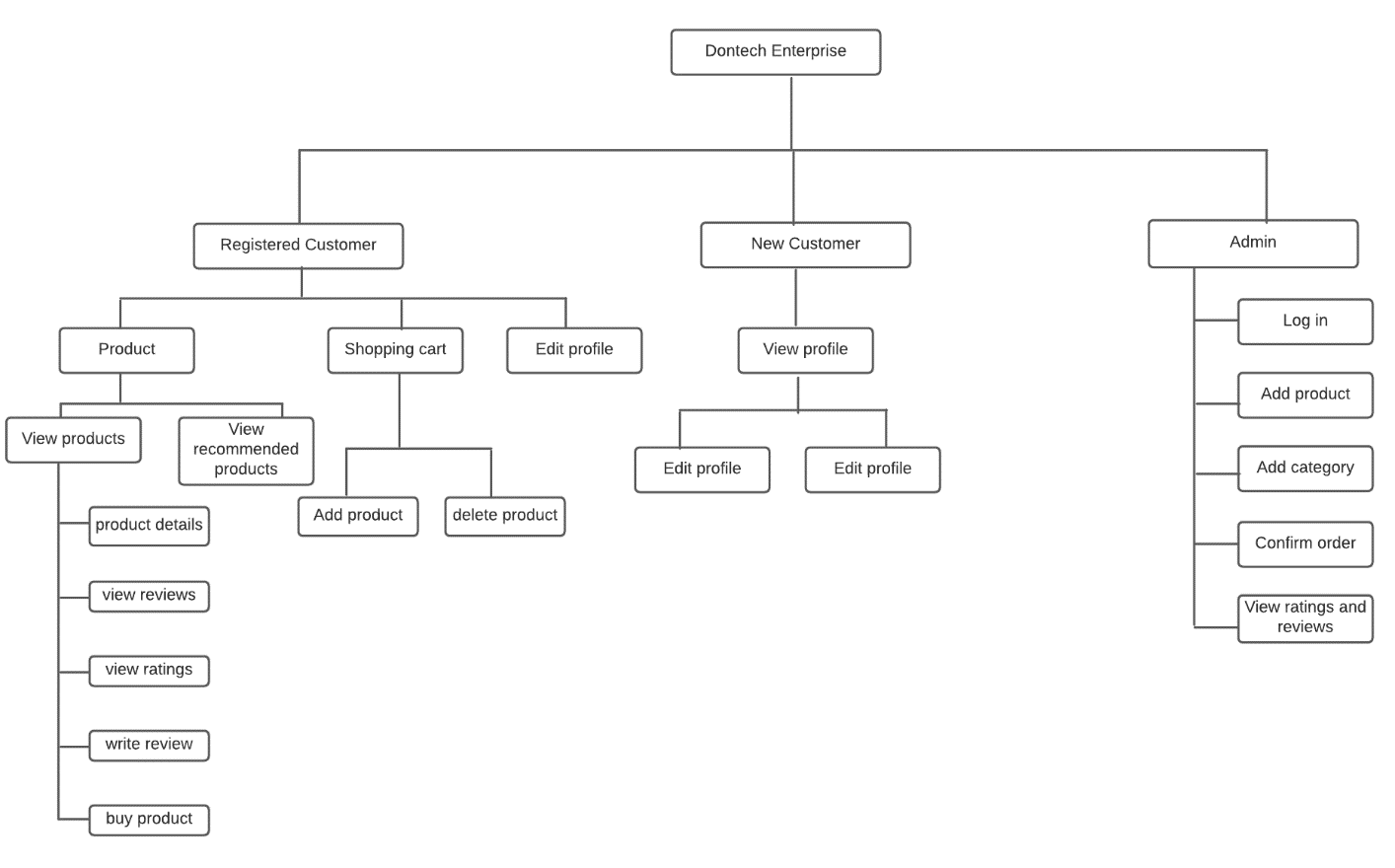


Figure Functional Decomposition Diagram

### 3.10.2 Use case diagram

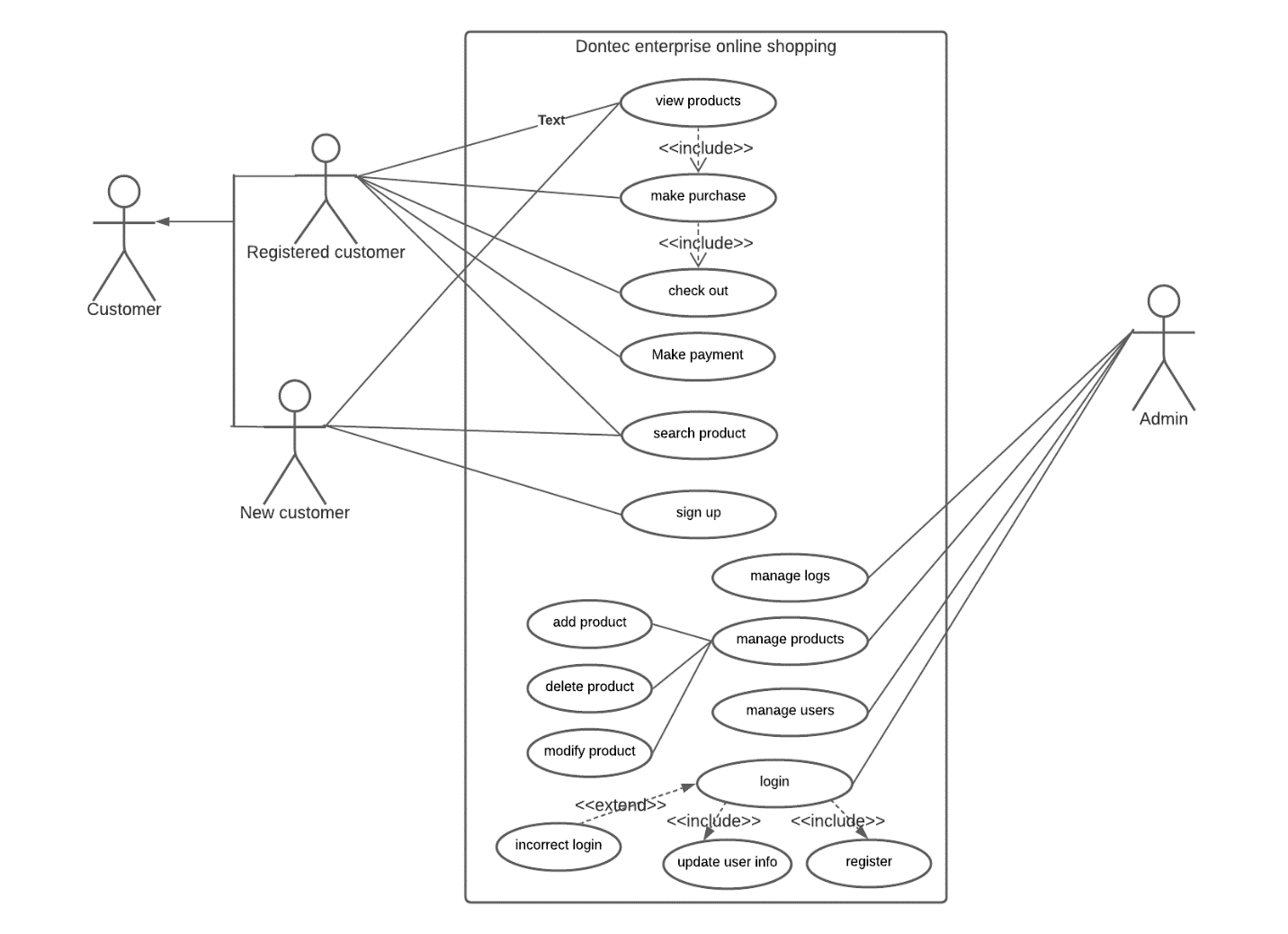


Figure Use Case Diagram

### 3.10.3 class diagram.

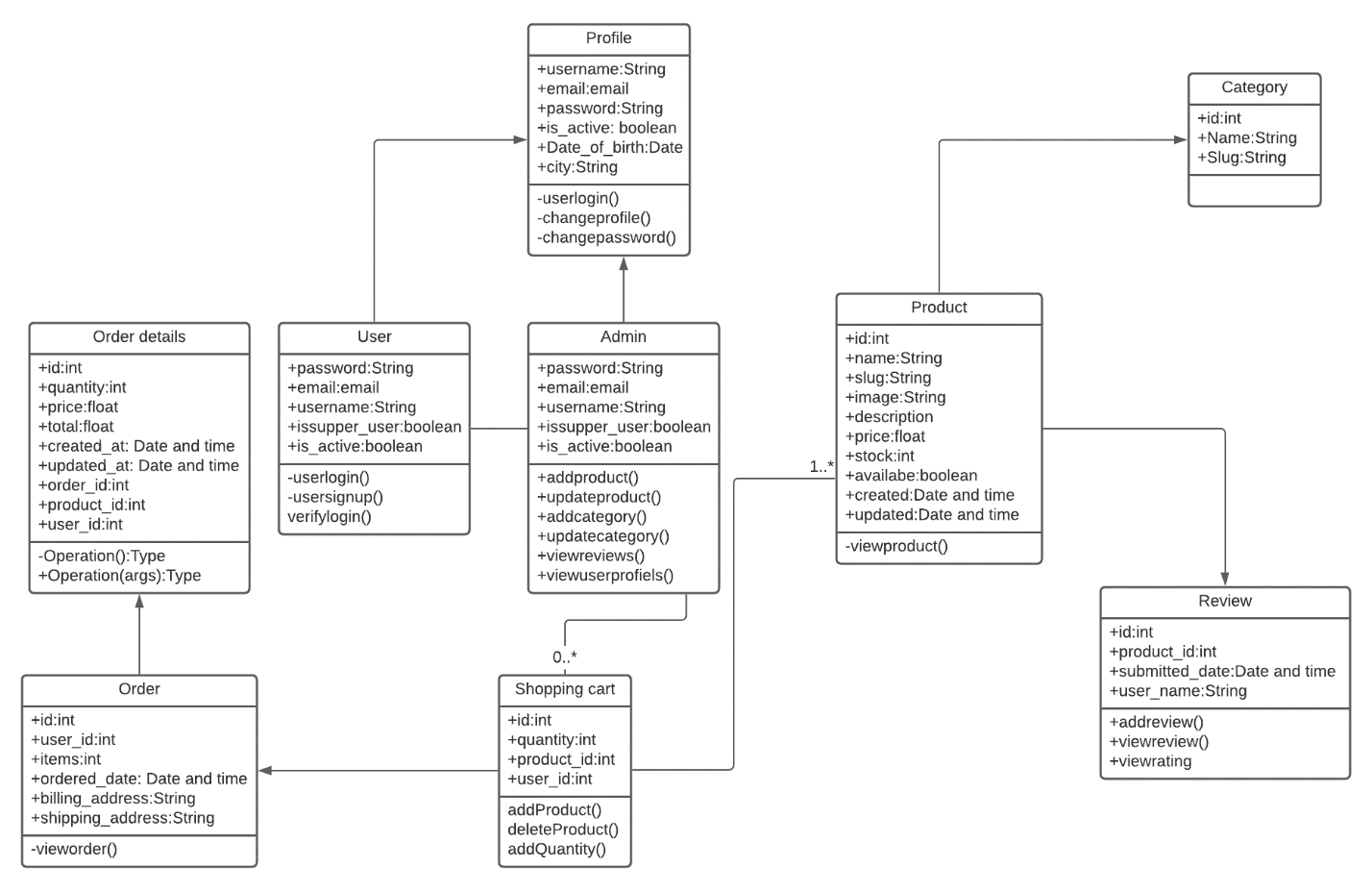


Figure Class Diagram

### 3.10.4 Activity diagram:

#### 3.10.4.1 Activity diagram for customers.

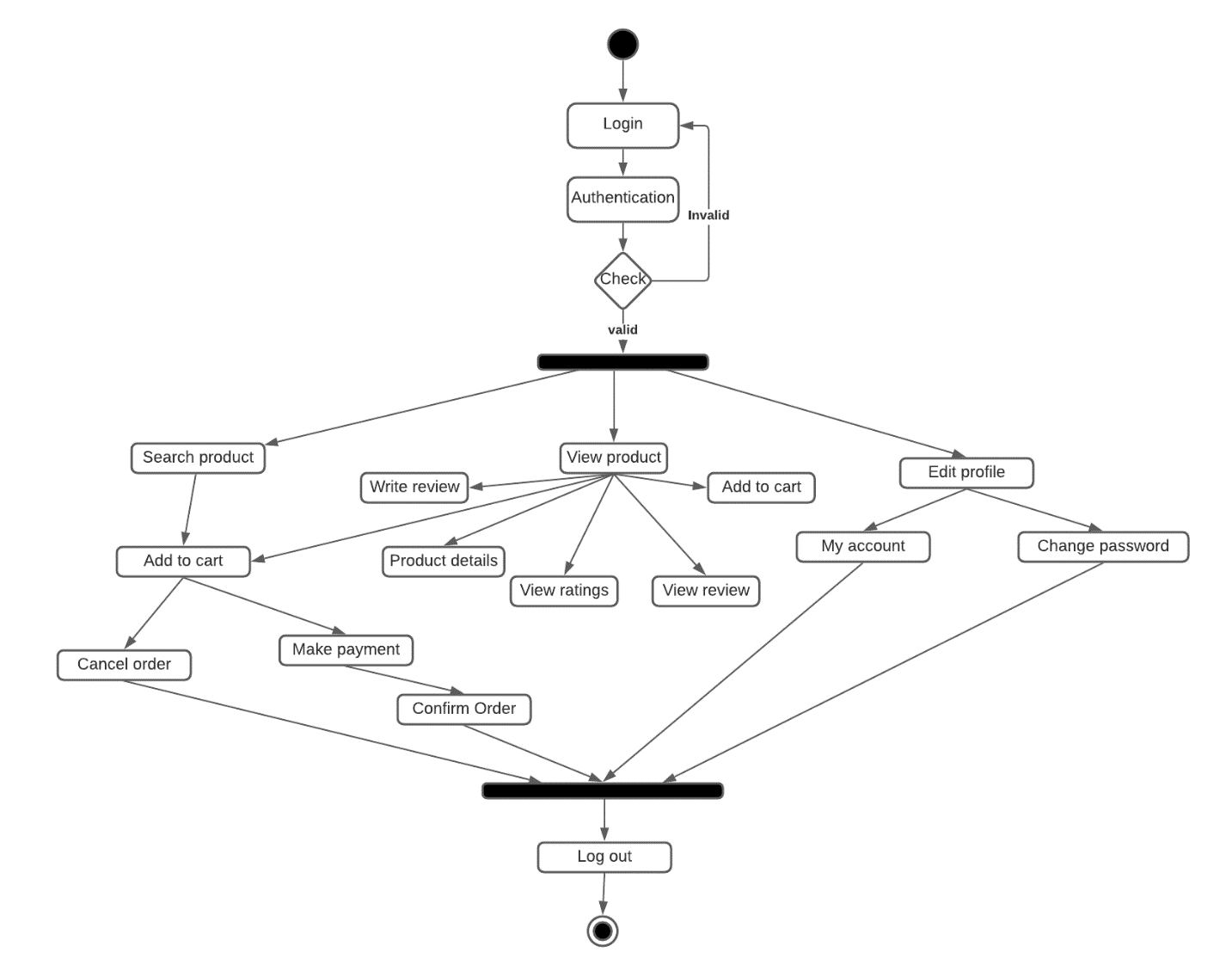


Figure Activity Diagram for Users

#### 3.10.4.2 Activity diagram for admin.

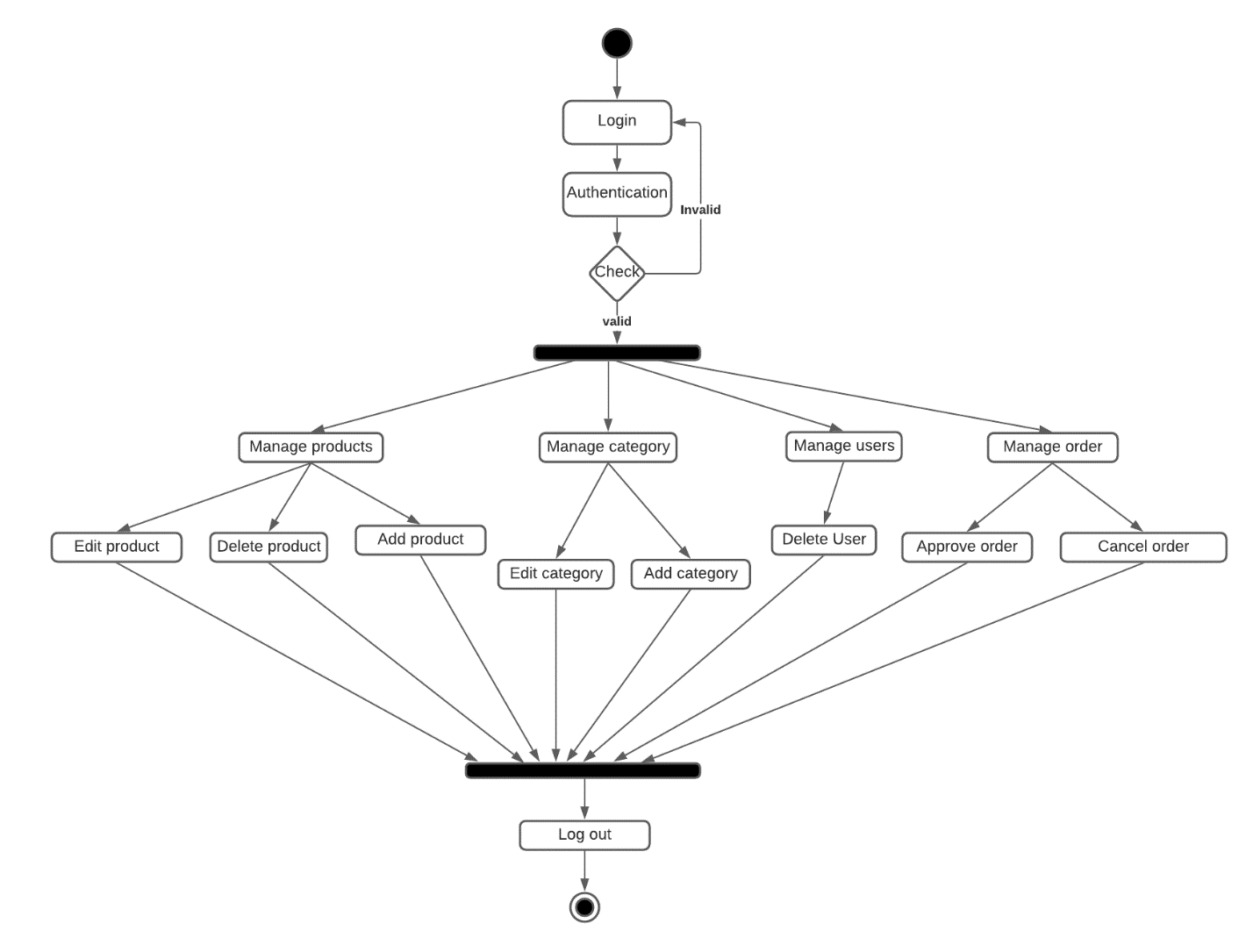


Figure Activity Diagram for Admin

# REFERENCES.

Alyoubi, A. A. (2015). E-commerce in developing countries and how to develop them during the introduction of modern systems. Procedia Computer Science, 65, 479-483.

Anantha, N. L., & Bhattula, B. P. (2017). A REVIEW ON RECOMMENDATION SYSTEM USING RATING DATASET. International Journal of Pure and Applied Mathematics, 116(5), 133-138.

Beel, J., Breitinger, C., & Langer, S. (2017). Evaluating the CC-IDF citation-weighting scheme: how effectively can ‘Inverse Document Frequency’(IDF) be applied to references. Proceedings of the 12th conference.

Bloem, C. (84). Percent of People Trust Online Reviews As Much As Friends.

Camilleri, A. R. (2020). The importance of online reviews depends on when they are presented. Decision Support Systems, 133, 113307.

Cassinelli, A., & Chen, C. W. (2009). CS224N Final Project Boost up! Sentiment categorization with machine learning techniques.

Chen, L., Chen, G., & Wang, F. (2015). Recommender systems based on user reviews: the state of the art. User Modeling and User-Adapted Interaction, 25(2), 99-154.

Gkatzioura, A., Sànchez-Marrè, M., & Gibert, K. (2018). A hybrid recommender system for industrial symbiotic networks. In iEMSs 2018 proceedings (pp. 1-6).

Gretze, U., & Yoo, K. Y. (2008). Use and Impact of online travel review: In the Proceedings of International Conference on Information and Communication Technology.

Harvie, C., Narjoko, D., & Oum, S. (2013). Small and Medium Enterprises’ access to finance: evidence from selected Asian economies. ERIA Discussion Paper Series, 23.

Hunaiti, Z., Masa’deh, R. E., Mansour, M., & Al-Nawafleh, A. (2009). Electronic commerce adoption barriers in small and medium-sized enterprises (SMEs) in developing countries: the case of Libya. IBIMA Business Review, 2(5), 37-45.

Hwangbo, H., Kim, Y. S., & Cha, K. J. (2018). Recommendation system development for fashion retail e-commerce. Electronic Commerce Research and Applications, 28, 94-101.

Kartiwi, M., & MacGregor, R. C. (2007). Electronic commerce adoption barriers in small to medium-sized enterprises (SMEs) in developed and developing countries: A cross-country comparison. Journal of Electronic Commerce in Organizations (JECO), 5(3), 35-51.

Katarya, R., & Verma, O. P. (2016, April). Effective collaborative movie recommender system using asymmetric user similarity and matrix factorization. In 2016 International Conference on Computing, Communication and Automation (ICCCA) (pp. 71-75). IEEE.

Kaur, A., & Monga, S. (2021). Digital Transformation in Business Era. In Managerial Issues in Digital Transformation of Global Modern Corporations (pp. 1-15). IGI Global.

Kim, H., Kim, J., & Herlocker, J. (2004, November). Feature-based prediction of unknown preferences for nearest-neighbor collaborative filtering. In Fourth IEEE International Conference on Data Mining (ICDM'04) (pp. 435-438). IEEE.

Kinuthia, J. N., & Akinnusi, D. M. (2014). The magnitude of barriers facing e-commerce businesses in Kenya. Journal of Internet and Information Systems, 4(1), 12-27.

Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. Educational and psychological measurement, 30(3), 607-610**.**

McLaughlin, M. R., & Herlocker, J. L. (2004, July). A collaborative filtering algorithm and evaluation metric that accurately model the user experience. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 329-336).

Mugenda, A. & Mugenda, O. (1999). Research methods: quantitative and qualitative approaches. Nairobi: Acts Press.

Ochola, F. O. (2013). Determinants of business collaterals and loan portfolio quality of commercial banks’ branches in Kisumu Municipality, Kenya (Doctoral dissertation).

Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010, pp. 1320-1326).

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. arXiv preprint cs/0506075.

Perset, K. (2010). The economic and social role of Internet intermediaries.

Shemi, A. P., & Procter, C. (2018). E-commerce and entrepreneurship in SMEs: Case of myBot. Journal of Small Business and Enterprise Development.

Sitikhu, P., Pahi, K., Thapa, P., & Shakya, S. (2019, November). A comparison of semantic similarity methods for maximum human interpretability. In 2019 artificial intelligence for transforming business and society (AITB) (Vol. 1, pp. 1-4). IEEE.

Sondur, M. S. D., Chigadani, M. A. P., & Nayak, S. (2016). Similarity measures for recommender systems: a comparative study. Journal for Research, 2(3).

Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: a survey. International Journal, 2(6), 282-292.

Vinodhini, G., & Chandrasekaran, R. M. EFFECT OF SAMPLING METHODS ON HYBRID CLASSIFICATION MODEL FOR SENTIMENT MINING.

Wanjau, K. L., Gakure, R. W., Magutu, P. O., & Kahiri, J. (2013). The Role Of Quality Adoptnion In Growth And Management Of Small & Medium Enterprises In Kenya.

Webster, J., Jung, S., & Herlocker, J. (2004). Collaborative filtering: a new approach to searching digital libraries. New Review of Information Networking, 10(2), 177-191.