INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration With

UNIVERSITY OF WESTMINSTER

Rankrr - Sentiment Analysis Based Product Ranking With Emphasized Informal Words & Emojis Interpretation

A Final Report by

Mr. E.M.D.R. Bandara

Supervised By:

Ms. Divya Premanantha

May 2023

This document is submitted in partial fulfilment of the requirements for the BEng (Hons) Software Engineering degree at the University of Westminster.

ABSTRACT

When it comes to recommendation systems which rank products through sentiment analysis, the sentiment scores of emojis are not considered. Also, emphasized informal words such as ‘liiiike’, ‘puuuuurfct’, or ‘yasssss’ are very common in user generated text contents in various online platforms. However, these words cannot be preprocessed properly with existing methods and will lead to poor correctness on tasks such as sentiment analysis which is a part of the NLP (Natural Language Processing) domain. These words are also not considered in sentiment analysis based recommendation systems which will affect in poor correctness on product rankings.

With the use of machine learning, the problem of emphasized informal words has been addressed by training a model based on a supervised classification algorithm with a synthesized dataset of emphasized words so that this model can predict the actual word which is meant by the emphasized counterpart. With this model as well as another model which considers emojis when analyzing sentiment, the correctness of sentiment based ranking process has been improved.

Tests on emphasized informal words model yielded good results on standard evaluation metrics. It also performs well on words with intentional misspellings with repeated letters while the ranking process was also improved according to the benchmarks.

Keywords - Recommendation Systems, Product Ranking, NLP, Sentiment Analysis, Machine Learning, Informal Text Preprocessing, Data Science

DECLARATION

I hereby declare that this research document and all the related artifacts are produced on my own and it has not been submitted before or is not currently being submitted for any degree program.

Full Name: Ekanayake Mudiyanselage Disura Randunu Bandara

IIT ID: 2018008

UoW ID: W1761865

10/5/2022

Signature Date

ACKNOWLEDGEMENT

I was able make this research project successful only because of the help and support I received from numerous people which had been a part of my life so far.

First, I would like to thank my supervisor, Miss Divya Premanantha for pushing me in the right way from the ideation stage of the research. She also suggested numerous ways to improve certain aspects of the research documentation as well as the actual implementation. Secondly, I would like to thank Mr. Guhanathan Poravi sir because he has been provided with the required knowledge not only for the final year research project but for several other modules as well, which were tremendously helpful throughout the research. I want to express my gratitude for the evaluators Mr. Tharindu Gimhana and Nayana Weerasekara for their opinion as well as the knowledge I gathered from them.

I would also like to thank the lecturers, management team as well as non-academic staff of IIT for the guidance and support I received throughout the years at IIT. I want to thank my friends at IIT as they have been immensely helpful throughout these past years. Finally, I want to express my thanks to my parents, siblings, and cousins for their support in every aspect of my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

NLP - Natural Language Processing

SA - Sentiment Analysis

RS - Recommendation Systems

ML - Machine Learning

1. INTRODUCTION

1.1 Chapter Overview

Product ranking is a technique to guide consumers to choose the right product online, as it becomes an essential feature for successful e-commerce or m-commerce applications. A well-functioning recommendation system can convert shoppers into buyers. Current recommendation systems mostly use star rating systems and textual reviews to rank items, but in social media, both emojis and text content are crucial in expressing a person's opinion. The main objective of this research is to build up a better product ranking system based on the sentiment of customer reviews which has emojis and text with the inclusion of informal text as well.

1.2. Problem Background

1.2.1. Product Ranking In Recommendation Systems

Customers prefer product recommendations that are tailored based on their interests. For example, when a customer feels indecisive about a product, people usually check out the "You May Also Like" section or "Similar Items" section.

Current recommendation systems mainly use content-based, collaborative filtering, or hybrid techniques to produce recommendations (Das, Sahoo and Datta, 2017). Product reviews which are written by users may include their preferences on certain product aspects and valuable explanations which are enriched with information about that product. Most of this information are ignored in recommendation systems. But this information highly matters when making a final decision to purchase a product.

1.2.2. Sentiment Analysis For Recommendation Systems

Exploiting customer experiences captured through reviews will be very beneficial to improve the recommendation systems. There are several works existing which analyze user reviews to improve this process. Opinion mining plays a huge role in this task as it's helpful to identify review elements such as discussed topic, the nature of opinion, context information, and emotions (Dang, Moreno-García and Prieta, 2021).

1.2.3. Emphasized Informal Words & Emojis Interpretation In Sentiment Analysis

Sentiment analysis is an NLP technique used to evaluate text content and classify the attitude intended by it. Usually, it will be classified into three groups: positive, negative, and neutral.

It is a broadly applied technique to understand user emotions towards a specific topic at hand (Yoo and Rayz, 2021).

Emphasized informal words (such as ‘liiike’, ‘puuuuurfect’) by repeating characters and intentional spelling mistakes are very common with users (Birjali, Kasri and Beni-Hssane, 2021). These words highlight the word and add more sentimental value to the text.

Emojis help users to convey their reactions and opinions more accurately than text. And they can be used to enhance sentiment analysis algorithms because emojis perform a very important role in expressing a user’s emotions (Wankhade, Rao and Kulkarni, 2022).

Properly analyzing the sentiment of reviews while considering emojis and emphasized informal text can improve the overall sentiment regarding a particular product which may affect the ranking process in recommendation engines to rank products more correctly.

1.3. Problem Definition

The problem of online retailing nowadays is that, as there are many products, their details and reviews on the internet are referred to as information overloading. Online retailers need to maintain a company reputation by providing quality services to the customers. If the product recommendations appear to the user at first are unacceptable products according to customers, they will move to another online seller to find out the product. This will reduce the sales of a particular online business and it will make it difficult for the business to survive for long in the online market. This is especially true in social media platforms as they have become much more popular in the e-commerce section. Also, when a recommendation system considers sentiments of customer reviews, the system needs to clean informal text as well as consider visual data such as emojis in order to analyze all the available data to improve the recommendations.

1.3.1. Problem Statement

The recommendation systems which are based on sentiment analysis don't consider emoji content when evaluating a customer review. Also, current sentiment analysis algorithms and preprocessing techniques won’t yield correct results on emphasized informal words with repeated letters or intentional misspellings (such as “liiiike” or “puuuurfect”), which may result in poor correctness.

1.4. Research Motivation

In this digital era, people use recommendation systems during various activities such as to find out movies or music, purchase products through online platforms and also on finding a partner on dating apps. Among various recommendation systems, product recommendation systems are one of the most widespread and mostly researched applications of machine learning in the field of ecommerce. By personalizing product recommendations to the users' preferences, it enables a better online shopping experience for the users. This is important for online businesses if they want to succeed in the digital platforms.

1.5. Research Gap

- Assessment of previous studies conducted on sentiment-based ranking in recommendation systems indicates that they don't consider the emoji content of user reviews when analyzing the sentiment score. Only the text content is evaluated when analyzing the sentiment score of an online review while emoji content is not taken into consideration (Heidary Dahooie et al., 2021).
- Also, the emphasized informal words always fallen into neutral polarity in current sentiment analysis algorithms which results in poor correctness because of inadequate preprocessing that fails to clean these noisy data from text (Birjali, Kasri and Beni-Hssane, 2021).

This can affect the ranking of products since there might be user reviews with only emojis or both text and emojis or when they are combined with emphasized informal words.

1.6. Contribution To The Body Of Knowledge

1.6.1. Contribution To Research Domain

Preprocessing Of Emphasized Informal Text - NLP

Current text preprocessing techniques won't yield desired results when cleaning the noisy text data which contains emphasized informal words. This will have a poor correctness when analyzing sentiment of a text which contains this type of data. This research will provide a machine learning based preprocessing technique where these texts are correctly cleaned which will result in improved sentiment analysis as well as improved ranking of products in a recommendation system since these texts carries valuable sentiment information.

1.6.2. Contribution To Problem Domain

Emoji Consideration When Ranking Products - Recommendation Systems

Today, almost every online user uses emojis to express their opinions. Emojis help customers more accurately express their emotions and thoughts. If a sentiment-based recommendation system can consider both emoji and textual content, which may also include emphasized informal text, it will improve the overall results of a recommendation system. Any improvement for the product recommendation systems will be advantageous to enterprises, manufacturers as well as end customers.

Preprocessing Of Emphasized Informal Text - NLP

Also, the proper preprocessing of emphasized informal words will be beneficial for NLP domain itself since these texts shifts the end result of the sentiment analysis task by analyzing them correctly instead of just categorizing them as text with neutral opinion when not preprocessed correctly.

1.7. Research Challenge

Emoji Consideration When Ranking

Current research doesn't consider emoji content when ranking the products through sentiment analysis in recommendation systems. This research considers implementing a method which considers the emoji content of customer reviews for product ranking. When developing the system, the strengths and drawbacks of available methodologies need to be thoroughly researched.

Preprocessing Of Emphasized Informal Text

Also, the existing text preprocessing techniques for cleaning emphasized informal text won't perform as required which will result in poor sentiment analysis. In this research, the approach for this will use a machine learning based technique, in order to correctly clean these noisy data. However, a suitable dataset or an appropriate mechanism to do this task were not available in the existing works. Hence, a custom dataset should be synthesized and a machine learning base solution should be developed from scratch.

1.8. Research Questions

RQ1: How to improve product ranking in recommendation systems using sentiment analysis?

RQ2: How products can be ranked and recommended for a user with sentiment analysis based on emoji and emphasized informal text interpretation.

RQ3: How to preprocess emphasized informal text with repeated letters or intentional misspellings using machine learning.

1.9. Research Aim

The aim of this research is to design, develop, and evaluate a solution which considers the sentiment of emojis and text when ranking products while also being able to correctly preprocess emphasized informal text.

This research will aim to deliver a technique that can rank and recommend products through sentiment analysis based on user reviews. Product ranking and emphasized informal text preprocessing will be the primary emphasis of this study. A machine learning based solution will be designed and developed to correctly preprocess the emphasized informal text. Data science techniques will be thoroughly explored to make the most viable method.

The essential knowledge will be gained, system components will be designed and developed, and the system will be assessed to prove the suggested hypothesis. The developed system will be able to run in a local environment or a remote server for public access. The source code, the used datasets, and the trained models will be accessible through a public repository for future research. A research document will be published about the result of the findings of the conducted research.

1.10. Research Objectives

Research Objectives	Description	Learning Outcomes	Research Questions
Literature Review	RO1: Identify existing methods and their limitations in product ranking and recommendation. RO2: Study how current recommendation systems incorporate sentiment analysis. RO3: Identify ways to interpret emojis and emphasize words in a text content. RO4: Conduct a study on how emoji and emphasized words interpretation can affect sentiment analysis.	LO1, LO4, LO6	RQ1, RQ2, RQ3
Requirement Elicitation	RO1: Gather requirements of product ranking techniques which are used in recommendation systems. RO2: Collect insights and ideas from technology and domain experts.	LO3	RQ1
Design	RO1: Improve the algorithm for estimating sentiment polarity based on the text content and the emojis in a review. RO2: Design a data pre-process mechanism to obtain a tokenized data set by separating text, emphasized text, and emoji content. RO3: Design a ranking method which can analyze text and emoji data to rank the related items.	LO2	RQ2, RQ3

	RO4: Design a machine learning based solution to preprocess emphasized informal texts.		
Implementation	RO1: Develop a ranking method which can analyze the sentiment of text, emphasized text and emoji content to improve the ranking of items. RO2: Develop a machine learning based solution to correctly preprocess emphasized informal texts.	LO7	RQ2, RQ3
Testing and Evaluation	RO1: Make an adequate test plan and perform relevant unit, integration, and functional tests. RO2: Evaluate how the developed ranking method improved over existing ones which doesn't make use of emoji content to rank the items. RO3: Evaluate how the machine learning based preprocessing technique cleans emphasized informal text in comparison to existing methods.	LO8	RQ1, RQ2, RQ3

Table 1: Research objectives

1.11. Chapter Summary

This section has covered the primary problem, research aims, research objectives, research gap and the contribution for the domains. This section summarizes the problem which the author is trying to address by developing a solution through research.

2. LITERATURE REVIEW

2.1. Chapter Overview

This chapter is about existing studies conducted on the field of sentiment analysis and product ranking based on sentiment analysis. This chapter will summarize the current techniques which are used in this field as well as the evaluation and benchmarking process.

2.2. Concept Map

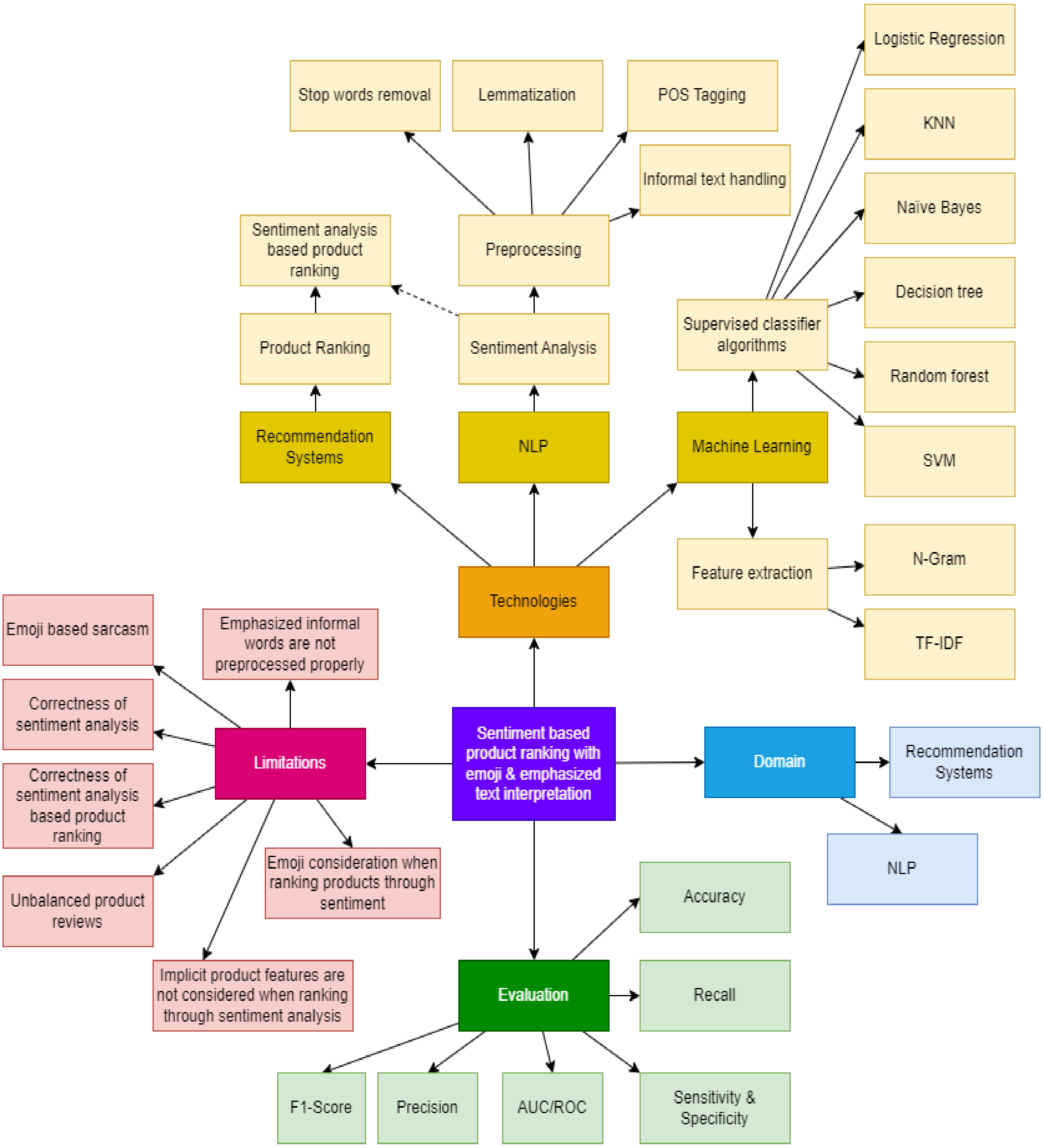


Figure 1: Concept map

2.3. Problem Domain

2.3.1 Recommendation Systems

Recommendation systems play a major role in various online applications. These systems are essentially helpers for users since these systems' primary functionality is about helping a user to find a product or a service by making suggestions based on their interests. This is done by creating relationships between the products, users, and their interactions to select the most suitable product for that particular user (Gasmi, Bouhadada and Benmachiche, 2020).

According to (Das, Sahoo and Datta, 2017), among the hierarchy of primary recommendation systems types, the personalized recommendations systems takes higher priority since it's based on individual user's interests.

Recommendation System Types

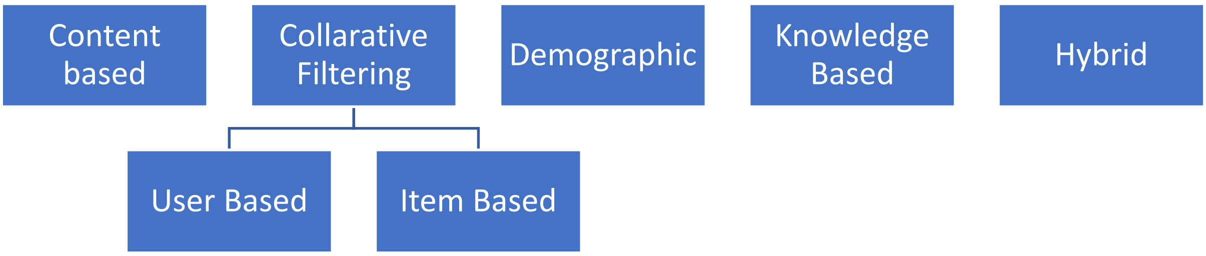


Figure 2: Personalized recommendation system types

Personalized recommendations systems are further divided into several categories as above based on their approach. Based on user's past purchases and present interactions, the content-based approach suggests items based on user's preferences and item descriptions. Collaborative filtering is about taking multiple users past purchases or interactions into consideration while suggesting an item that the user might purchase. Basically, it's about filtering products based on users with similar interests. Demographic filtering considers user's age, gender, and other factors such as location to suggest products. Knowledge based recommendation systems came into the market to address the problem of products having a small history of purchases. User preferences are taken explicitly and then consider the product and its features to suggest for the user. Hybrid systems are about combining these different approaches to make a better system to provide suggestions for users. Disadvantages of one approach can be reduced by another approach in hybrid systems to build a robust system (Das, Sahoo and Datta, 2017).

Recommendation System Feedback Techniques

Suggestions are based on information as it gives the data for the recommendation system in order to make suggestions for users based on their preferences (Das, Sahoo and Datta, 2017).

- **Implicit Feedback**

  This method is about extracting information from the users without their consciousness. All the actions that are performed are taken in as information in this approach. Mouse events and screen time are few examples of these actions. In this way, the user's interests are captured without their consent in the background.

- **Explicit feedback**

  This method requires the users to give a rating or a review for the product to evaluate it. With these rating measurements there is a numerical value to statistically provide recommendations based on metrics such as the average of the ratings.

There is also the hybrid feedback technique which combines numerical ratings with user interactions to get information on user interests to further improve the recommendation algorithm.

This research is based on explicit feedback technique since it's about extracting reviews in order to give a sentiment score which will be used as a numerical rating to sort and provide suggestions based on a statistical value such as average of the sentiment score.

2.3.2. Sentiment Analysis

According to (Tejwani, 2014), sentiment analysis (SA) is a sub task of natural language processing (NLP) which aims to obtain subjective information by mining the opinion expressed by a human on a given material. This extracted opinion can be used to make better decisions and gain insights on various domains.

Sentiment analysis is also a key component in the development of artificial intelligence as well. It can be regarded as a classification task since it's about performing several tasks which categorize a given content into either positive, negative, or neutral based on how these contents are expressed by humans. This process also involves NLP sub tasks such as subjectivity and sarcasm detection (Birjali, Kasri and Beni-Hssane, 2021).

Levels of Sentiment Analysis

- **Aspect Level**

  This level is about analyzing sentiment on aspects or features of a given entity to perform fine grained analysis to gain more details about aspect opinions produced by users. These aspects can be implicit or explicit depending on the situation and the needs (Wankhade, Rao and Kulkarni, 2022).

- **Sentence Level**

  Sentience level analysis is about focusing on each sentence of a given text to gain the sentiments of these individually.

- **Document Level**

  This level aims to perform the SA task on the whole text to determine what opinion it expresses.

Out of these multiple levels, aspect level sentiment analysis is very important for recommendation systems since it can be used to extract opinions on product features which can be used to provide better recommendations.

Sentiment Analysis Pre-Processing

To perform the tasks of SA accurately, several other NLP preprocessing tasks need be combined. Data from various online resources such as social media or ecommerce platforms contains unstructured data with lot of noise in them which includes spelling or grammar mistakes as well as informal texts (Birjali, Kasri and Beni-Hssane, 2021).

- **Tokenization -** This split the given texts into smaller items to be processed.
- **Stop Words Removal -** Words such as ‘an’, ‘the’, ‘or’, ‘for’ are removed since they don’t hold any sentimental value.
- **POS (Part of Speech) Tagging** - This step categorizes and tags the tokenized items into verbs, adjectives, nouns, and adverbs.
- **Lemmatization** – This step converts a word into the base form. (Ex: ‘Beautifully’ into ‘Beautiful’)

Pre-Processing of Emphasized Informal Words

However, some text content requires more cleaning tasks in order to properly remove the noise in them. Words such as 'liiiiiike' or 'puuuuuuurfct' are not grammatically correct. But these words are intentionally entered by the users and also contain valuable sentimental information (Birjali, Kasri and Beni-Hssane, 2021). Also, words with '*' characater in between them represent swear words which may hold negative sentimental information. These types of informal words are very common in online platforms, especially in social media and ecommerce platforms.

Emoji Interpretation In Sentiment Analysis

In almost any online platform, users incorporate emojis to express their emotions more clearly and visually. The usage of emojis has been growing significantly over the recent years. Sentiment analysis can be improved with the use of emojis since they provide sentimental value. However, there are existing problems on the actual impact of emojis in certain text which may represent sarcasm. There are also studies on replacing emojis with a set of related words to improve the SA. Nonetheless, emojis plays a huge role in the field of NLP and can be used as a valuable data source to improve the NLP tasks such as SA (Yoo and Rayz, 2021).

Machine Learning For Sentiment Analysis

After proper data preprocessing, the data can be fed into a machine learning model to train the classification of sentiment polarity. Machine learning approaches can be used to get better sentiment analysis results based on the patterns of the text. However, a good dataset is needed in order to achieve better performance in this task.

2.3.3. Sentiment Analysis For Recommendation Systems

Users generate massive volumes of product reviews in various online platforms including ecommerce and social media. These reviews provide valuable feedback for businesses to improve their products as well as helping other users to make decisions on their purchases. In the field of ecommerce, the user opinions on key attributes of a particular product shape the brand's image. These user reviews provide an excellent approach to mine the opinion of a product in the online market. A user's purchase decisions can be improved based on this opinion provided by past experiences and purchases of another user. With the help of sentiment analysis, a recommendation system can be implemented which works on the users' opinion on products. Generally, it's done by classifying the reviews into the sentimental categories which are namely neutral, negative, and positive. The concept of aspect level sentiment analysis provides a fine-tuned sentiment analysis on a product since it analyses the features of products and the reviews which contains text content about these features (K and S, 2022).

With the help of sentiment analysis on product recommendation systems, users can get better recommendations based on opinions from other users about the products to make better decisions on purchases to improve the overall online shopping experience. Solving the issue of handling emphasized informal words and the interpretation of emojis will remarkably improve the sentiment values for a given text content. Hence, this can be applied to recommendation systems to improve their overall correctness so that they can provide better suggestions for users.

2.4. Existing Work

2.4.1. Sentiment Analysis Based Product Ranking

Including sentiment analysis in recommendation systems can be very beneficial. Several studies have been conducted on sentiment-based product ranking and it's keep getting improved. However, these existing studies won't consider the emojis when analyzing the sentiment scores.

(Zhang, Wu and Liu, 2020) suggest an approach which analyzes sentiment score for each product feature in every review of that product which can be used to calculate the overall score of each feature on difference products based on hesitant fuzzy set theory which can be used when there are indications of hesitations on features among the reviews.

The study conducted by (Heidary Dahooie et al., 2021), addresses several gaps by combining sentiment analysis with multi criteria decision making mechanisms along with intuitionistic fuzzy sets. Focusing on specified product features or selecting them through the occurrence frequency is a primary problem which is addressed by this approach. Indications of hesitations in reviews and proper feature weight calculation has been incorporated to provide a better algorithm.

The approach by (Bi, Liu and Fan, 2019a) uses interval type-2 fuzzy sets to represent sentiment analysis results to handle uncertainty indicated in reviews. This way, results with an accuracy of 100% score and limited amount of score will be used to gain sentiment analysis data on reviews which can be used to improve the product ranking.

Another approach is studied by (Wu and Zhang, 2019), which uses IF (Intuitionistic Fuzzy) theory to calculate sentiment scores for each product feature for all the products, which will then be used with multi criteria decision making methods to rank products.

The study conducted by (Guo, Du and Kou, 2018), is based on the combination of objective and subjective sentiment values of product features to get new sentimental scores. Product feature weights are determined with LDA (Latent Dirichlet Allocation) topic model to get the objective sentiment scores. These scores are taken into consideration along with user preferences which will be used to rank the products through 'PageRank' algorithm.

Emoji Interpretation For Sentiment Analysis Based Product Ranking

Even if there were numerous research on sentiment analysis based product ranking, there was no consideration of emoji data in these studies. However, emojis hold valuable sentiment information as mentioned by (Yoo and Rayz, 2021) which can be of impact when used for sentiment analysis based product ranking techniques.

2.4.2. Emphasized Informal Text Pre-Processing For Sentiment Analysis

Informal words are very common in user generated text content in online platforms. Especially emphasized texts since these kinds of words either intentionally misspelled or include repeated letters to highlight their meaning to express a strong emotion from them. Existing studies on this problem are very rare and overlooked.

(Birjali, Kasri and Beni-Hssane, 2021) states that these words with repeated letters can be very noisy and therefore they need to be preprocessed to in order to achieve better results on sentiment analysis.

Another method is reviewed by (Giachanou and Crestani, 2017), which suggest that feature selection can be used to solve the problem of having words with repeated letters. However, there was no clear approach to this method.

(Agarwal et al., 2011) suggests to preprocess emphasized informal words by replacing the repeated letters by a sequence of three repeating letters (Ex: 'coooool' to 'coool'). This may work on very specific occasions if a sentiment analysis model was built on this particular set of preprocessed data. However, it may fail to perform on new and real time data since there can be numerous variations on words with repeating letters. Also, this approach won't be able to consider words with intentionally misspelled words with repeating letters. This would be another issue in real time data.

Spelling Correction Approach

The approach proposed by (Li et al., 2018), which is based on a nested RNN (Recurrent Neural Network) model that tries to solve the spelling correction by combining two RNN models which analyses characters and words to predict the correct word out of the misspelled one. This may work on few occasions of having repeated letters in a word, but it may not be able to accurately correct the misspellings of intentionally misspelled words. Because in this approach, one model goes through each letter and uses another model to predict the correct word after analyzing the letters. This approach won't work for non-existing words with intentionally misspelled words since there aren't any actual words to predict.

(Nuspell, no date) is an open-source spell checker program which is used by popular organizations such as Mozilla and Ubuntu, to build their own spell checker programs in their applications on top of it. However, these approaches won't work on emphasized informal text and fail to provide correct spellings or the actual words which are meant by them.

There is also the possibility of removing repeated letters to get the actual word. Another possibility is that after this task, a dictionary can be used to get the nearest word which matches the partially cleaned emphasized word. However, there are situations where this won't work as there are numerous variations of emphasized words combined with repeated letters and intentional misspellings.

2.5. Technological Review

The primary component of this research is the emphasized informal text pre-processing component which will use machine learning. Since this research tries to solve this problem by a text classification technique, the overall process has been reviewed. Since this research tries to classify words instead of whole texts, only the most relevant sub tasks will be discussed.

2.5.1. Pre-Processing

Generally, in machine learning approaches the pre-processing task is important in order to clean the data and remove unwanted noise from them. This step prepares the dataset in the required format which will be then fed into a model to train upon them.

According to the survey study conducted by (Kowsari et al., 2019), the following steps are important for the task of word classification:

- **Capitalization -** Convert the words into lower case in order to remove inconsistencies with capitalization. This is a rule of thumb when it comes to
- **Duplications Removal -** There might be duplications among the dataset. These needs to be removed to get a proper dataset.

2.5.2. Feature Extraction

This step is required to identify semantic relations between the letters of a word in order to train a model on these features so that it can predict and classify a given input as discussed by (Kowsari et al., 2019).

- **N-Gram**
  A technique which can be used to represent the order of letters in a word that can be used to represent a feature.
- **Count Vectorization**
  Converts the word into a matrix of letter occurrences which can be used to extract other features as well.
- **TF-IDF (Term Frequency-Inverse Document Frequency)**
  This method assigns a higher weight to letters with high occurrences in a word. This can be used to identify repeated letters in the word and make it a feature.

2.5.3. Dimensionality Reduction

To address the issue of time complexity and resource consumption, dimensionality reduction can be used to reduce the size of feature space to effectively manage the resources (Kowsari et al., 2019).

2.5.4. Classification Algorithm

Since the primary task is to classify a given word, several classification algorithms were reviewed. It's not possible to conclude on a best classification algorithm since they depend on various aspects of a given problem.

When it comes to machine learning, supervised learning is one of the primary branches. The use of past data is a key factor in supervised techniques. After the training process, one can use that model to predict the outcome for new data. Since the data science component of this research is trying to address a problem which can be solved by categorizing, classification algorithms which are based on supervised learning techniques will be suitable (Sen, Hajra and Ghosh, 2020).

As discussed by (Kowsari et al., 2019), there are several machine learning based classification algorithms which can be used in this research. Also, some of the mostly used supervised algorithms are reviewed by (Sen, Hajra and Ghosh, 2020) which gives an overall overview of them. Since the algorithm is a vital part of the primary component, the available algorithms must be studied in order to select an adequate one for this research.

Classifier Algorithm	**Advantages**	**Disadvantages**
Multi Nominal Naïve Bayes	• Scales well with large datasets • Discrete feature handling with the use of feature distribution by multinominal distribution assumption.	• Assume that features are conditionally independent on classes. • Won't perform well for imbalance dataset.
Decision Tree	• Considers complex non-linear relationships among features and target variables. • Inherently perform feature selection on most useful	• Overfits the training data • Sensitive to small changes on the training dataset which will reduce the robustness

	features on every split.	
Random Forest	• High accuracy in classification tasks since it uses multiple decision trees. • Can handle a large number of features without dimensionality reduction techniques.	• Significant memory consumption for large models. • Imbalanced dataset may cause to be biased towards the majority class.
Support Vector Machine	• Effective on a smaller training dataset. • Can address the issue of imbalance dataset by giving more weight to minority class.	• Sensitive to hyper parameter tuning. Careful experimentation is required to achieve optimum performance. • Heavy on resource consumption and can be computationally expensive on large datasets.
K-Nearest Neighbor	• Can handle complex and non-linear relationships. • Can be used to multi-class classification problems	• Imbalance datasets can lead to biased results. • Need to standardize features because otherwise, may cause unfairness.

Table 2: Supervised learning classifier algorithms advantages and disadvantages

2.5.5. Hyper Parameter Tuning

Since there are multiple parameters it's best to use a method to get the best combination of parameters which results in the most performance. By doing so, most of the manual tuning process can be eliminated and make the training process much more efficient (Kowsari et al., 2019).

2.6. Evaluation and Benchmarking

2.6.1. Word Classification Model Evaluation

Since the primary data science component of this research is based on the text classification problem, several evaluation techniques were reviewed which were suggested by (Kowsari et al., 2019). Standard classification evaluation metrics were selected for this research as well.

- **Accuracy -** Represents the ratio of correct predictions against the total input dataset.
- **Precision -** Ratio between true positive predictions and all positive data.
- **Recall -** Measure the model on correctly predicting true positive data.
- **F1-Score -** The harmonic mean of precision and recall scores which can be used to evaluate the precision as well as its robustness.
- **Specificity -** Measures the ability to predict a true negative on every class.
- **Sensitivity -** Measures the ability to predict a true positive on every class.
- **AUC / ROC -** Combined measurement of performance on all classification levels.

2.6.2. Benchmarking

Most of the existing studies compare their proposed approach with other approaches to benchmark the systems. This method will be used in this research as well to benchmark against existing systems which perform similar tasks.

2.7. Chapter Summary

This section summarized the problem domain, existing works, currently used technologies, and evaluation with sufficient literature which relates to recommendation systems and text classification tasks. Through the literature review process, the existing research gaps were identified to address them in this research.

3. METHODOLOGY

3.1. Chapter Overview

This chapter describes the methodologies which will be carried out through different phases of the research. It also contains resource requirements and risks mitigation information.

3.2. Research Methodology

Research Philosophy	The proposed system will eventually evaluate the outcomes as positive, neutral, or negative and will rank these outcomes. To implement this, these outcomes will be interpreted with numerical values. Which means this is quantitative research. Hence, **Positivism** has been chosen as the research philosophy.
Research Approach	The **Deductive** approach has been chosen because this is quantitative research that aims to experiment and prove the hypothesis which was built to improve an existing theory.
Research Strategy	**Questionnaire, Experiments and Literature Review** were chosen as strategies. Questionnaire seems to be the primary strategy while Experiment strategy is also will be used since the research is trying to improve an existing theory by straightforward experimental process.
Research Choice	The **Multi-method** has been chosen as the research choice because the research is mainly about text and emoji content which belongs to the qualitative data category which will be interpreted as quantitative data to prove the hypothesis.dd
Time Horizons	The data will be gathered over time since the proposed system in this research is constantly analyzing and evaluating this data to improve the results. Due to this reason, **Longitudinal** was chosen as the time horizon.

Table 3: Research Methodologies

3.3. Development Methodology

3.3.1. Software Development Life Cycle Model

Since most of the specifics about the project is unknown at the time of starting, the development will be iterative. So, the **Agile** software development life cycle model has been chosen.

3.3.4. Development Methodology

The proposed system will contain various objects with encapsulated behavior and data. These objects will interact with other components of the application to perform the required tasks. Due to this, **Functional Programming** has been chosen as the development methodology.

3.4. Project Management Methodology

PRINCE2 Agile has been chosen as the project management methodology as it allows to manage the research project with flexibility in mind.

3.5. Resource Requirements

3.5.1. Hardware Requirements

- **CPU: 4 Core CPU or above** - To perform CPU intensive tasks and to train the data science model.
- **RAM: 16GB or above** - To run the required development tools and manage the datasets.
- **Disk Space: At least 20GB** - To store the relevant data for research.

3.5.2. Software Requirements

- **Operating System (Windows 10 with WSL) -** Windows 10 is the default selection for development due to ease of use. If a software tool requires a Linux environment, WSL can be used within Windows 10 to execute it.
- **Python -** The base programming language for the data science component and it's API.
- **Angular** - To develop the front-end application to demonstrate the behavior of the recommendation engine along with the improvement done through the research.
- **Anaconda** - To create the local development environment with relevant data science related technologies and libraries.
- **PyCharm** - To develop, train and test the machine learning models in the local environment.
- **VS Code** - As the Primary IDEs for code editing.
- **MS Office / Google Docs** - To create and maintain the documents.
- **Git and GitHub** - As the primary version control system and the remote source code repository.
- **Google Drive** To back up documents and other relevant files as needed.
- **Zotero** - Reference management software to collect, organize and maintain references.

3.5.3. Technical Skill Requirements

- **Data science skills -** To design, develop and evaluate the data science component.
- **UI/UX design skills -** To design and develop the user interface for the demonstration web app.

3.5.4. Data Requirements

- **Product reviews data -** Amazon reviews datasets from Google Dataset Search or Kaggle. If not available, will be manually collected through a script from a publicly available resource such as a social media platform.
- **Twitter Sentiment140 -** Kaggle.
- **Emphasized informal text data -** Manually synthesized.

3.5. Risks And Mitigation

Risk	Probability of Occurrence	Magnitude of the Loss	Mitigation Plan
Time will not be enough to gather required skills and knowledge	5	5	Proper time management by following a schedule.
If a proper dataset could not be found, it could obstruct the development and completion of the research.	5	5	Collect a proper dataset from publicly available sources.
Since this is a data science related research, an adequate knowledge about the domain is required to complete it.	5	5	Get advice, ideas and insights from the supervisor, lecturers, and colleagues to continuously improve the knowledge.
Developed source code or documentation can be corrupted or lost due to device failures.	4	5	Back up the source code on GitHub. Adhere to a cloud-based documentation approach. An external device can be used to backup as an extra

			precaution.
May not be able to stay updated with technologies and the domain since they are constantly updating and keeps getting evolved over time. This might cause the research to not being up to date.	4	5	Regularly check for technology and domain updates and discuss with the relevant professionals, supervisor, lecturers, and colleagues.

Table 4: Risks and mitigation

3.6. Chapter Summary

This chapter summarizes relevant methodologies which are going to be carried out throughout the research. The resource requirements are well described, and risks mitigation is also being considered and included in a tabular format.

4. SOFTWARE REQUIREMENTS SPECIFICATIONS

4.1. Chapter Overview

This section is about identifying possible stakeholders which will interact with the system in various aspects. It also includes the requirement gathering methodologies and the findings of the requirement gathering task which is used to identify the functional and non-function requirements for the prototype development.

4.2. Rich Picture Diagram

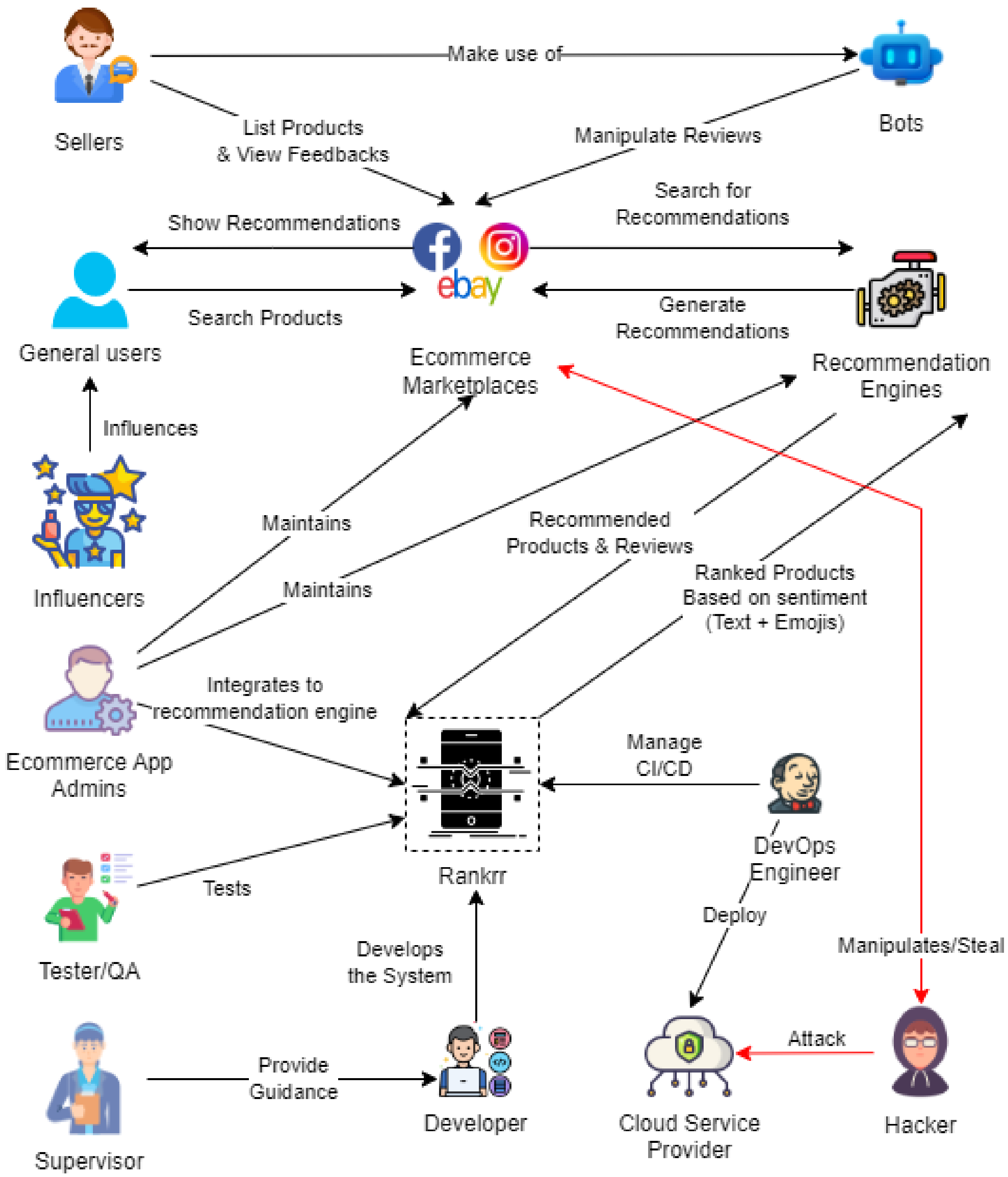


Figure 3: Rich picture diagram

4.3. Stakeholder Analysis

4.3.1. Stakeholder Onion Model

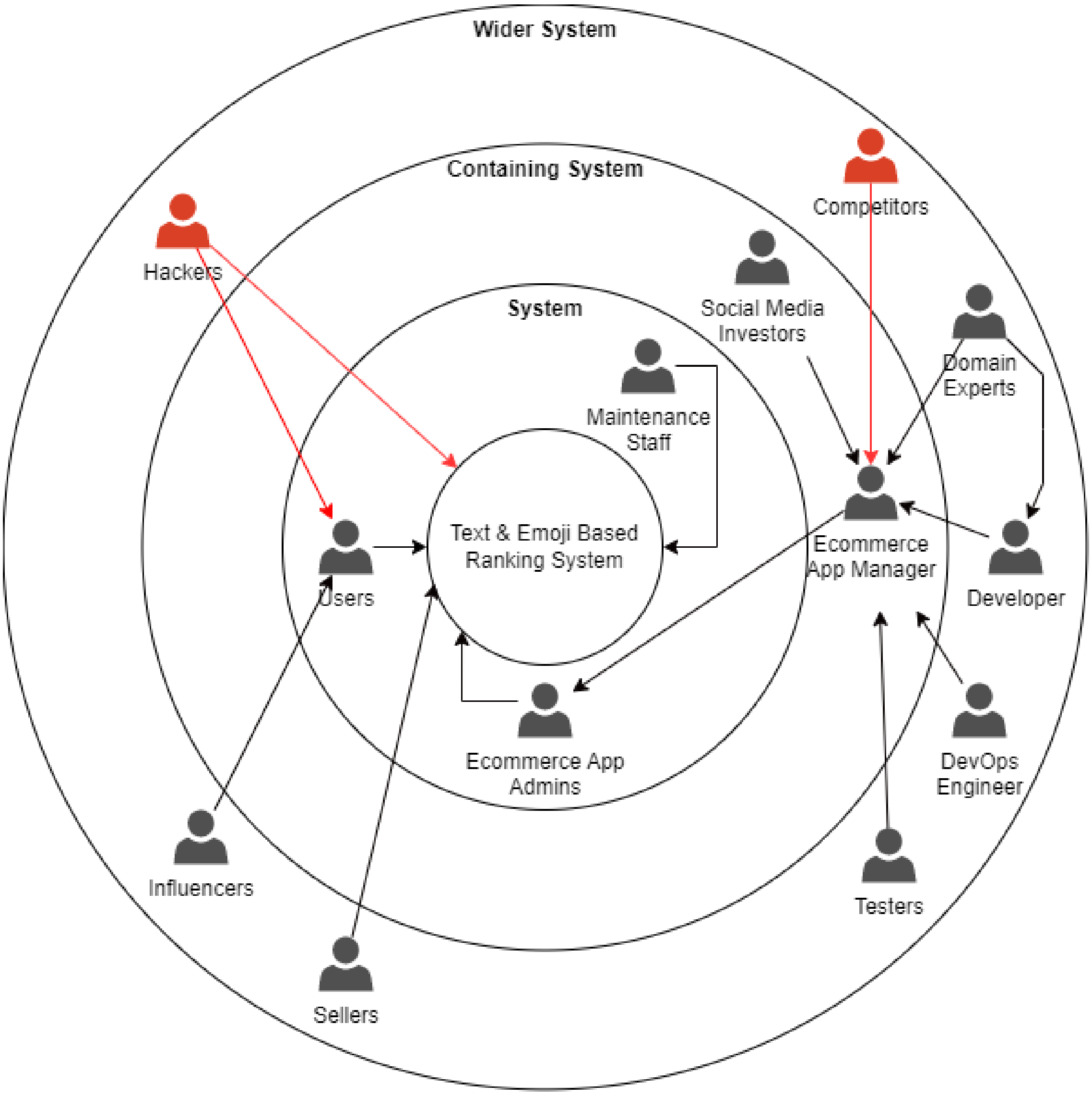


Figure 4: Stakeholder onion model

4.3.2. Stakeholder Viewpoints

Stakeholder	Role	Description
Developer	Developer, Financial Beneficiary	Develops the proposed system.
Domain Experts	Expert	Provide advice and suggest improvements for the system.
Sellers	Fundamental Beneficiary	Makes a profit when a recommended product has been purchased through a user.

Users	Fundamental Beneficiary	Get relevant product recommendations to purchase in the end.
Influencers	Secondary	Influence users and drive trends through social media platforms.
DevOps Engineers	Deployment & Maintenance	Deploy into a cloud service and maintains.
Maintenance Staff	Support	System maintenance
Testers	Quality Inspector	Test the system before and after deployment.
Ecommerce App Manager	System Owner	Oversees the system ecommerce admins and maintenance.
Ecommerce App Admin	Administration	Modifies the required parameters of the system, check status metrics, provide data sources, integrates with social media platform, and ensures the system is operational without issues.
Competitors	Negative Stakeholder	Develops similar systems which may surpass the developing system.
Hackers	Negative Stakeholder	Manipulates or steal data through social media for personal gains.

Table 5: Stakeholder viewpoints

4.4. Requirement Elicitation Methodologies

Methodology	Reason For Selection
Survey	A survey will be useful to determine certain requirement expectations for the proposed system. This method will also be useful to determine if the system will be useful for the end users. A questionnaire has been used to conduct the survey for this research.
Literature Review	Literature review process is very helpful to identify current solutions and their limitations. Through this process, various technologies and methodologies which are relevant to this research can be identified as well. A through literature review has been done and several research gaps were analyzed along with the current methodologies and the used technologies.

Table 6: Requirement elicitation methodologies justifications

4.5. Discussion Of Findings

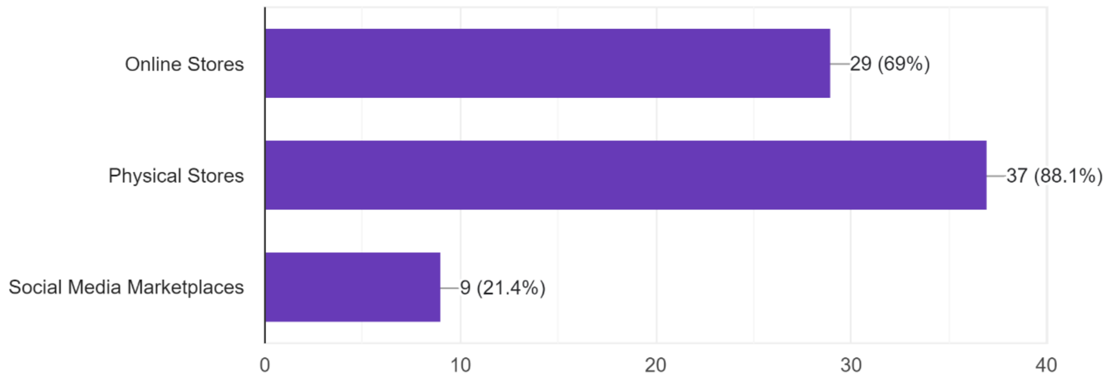
4.5.1. Literature Review

Finding	Citation
Most of the current systems do not consider the endogeneity of social media when generating recommendations. Influence can greatly affect the outcome of recommendations.	(Wang, Wang and Song, 2017)
Word sense disambiguation, sarcasm detection, emphasized word analyzing and negation handling are the most challenging tasks in sentiment analysis. This process becomes much more complex when emojis are mixed with text because they can greatly affect the actual meaning of the text.	(Birjali, Kasri and Beni-Hssane, 2021)
When it comes to product aspect level sentiment, only the explicit product aspects expressed by explicit words have been considered when ranking through sentiment analysis. Dynamic product aspects are not considered, and it can affect the overall sentiment value of the product.	(Heidary Dahooie et al., 2021)
Unbalanced review amounts can affect the opinion about certain products. In this case, the bias towards one product should be ignored to get a valid sentiment score.	(Tian et al., 2009)
User preferences were not taken into consideration when ranking the recommendations through sentiment analysis. This would negatively affect on user experience since there may be instances where users get irrelevant recommendations.	(Bi, Liu and Fan, 2019b)
A support system is needed for the end users to make use of the ranking mechanism. Also, new products with few reviews are not considered. This issue of having unbalanced and lack of data will provide inaccurate rankings.	(Liu, Bi and Fan, 2017)

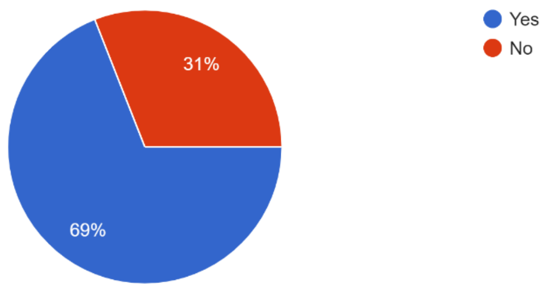
Table 7: Literature review findings

4.5.2. Survey

<table>
<tr><td>Question</td><td>What is your preferred method to purchase products?</td></tr>
<tr><td>Aim</td><td>To identify the most preferred platform for online purchases.</td></tr>
<tr><td colspan="2">Findings

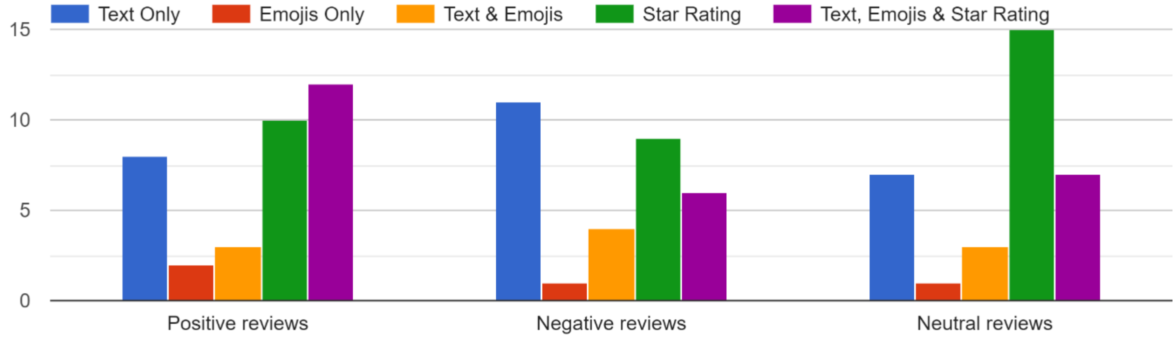


Mostly uses physical stores instead of online platforms to purchase products. However, online stores themselves has more preference over the social media marketplaces when it comes to online platforms. This may be due to the integrity of social media marketplaces.</td></tr>
<tr><td>Question</td><td>Have you ever reviewed any product?</td></tr>
<tr><td>Aim</td><td>To identify whether the participant has reviewed a product in an online platform after purchasing.</td></tr>
<tr><td colspan="2">Findings



Most of the responders has reviewed products in online platform.</td></tr>
<tr><td>Question</td><td>If yes, what is your typical way of reviewing a product?</td></tr>
<tr><td>Aim</td><td>To identify the typical way of reviewing a product after purchase.</td></tr>
</table>

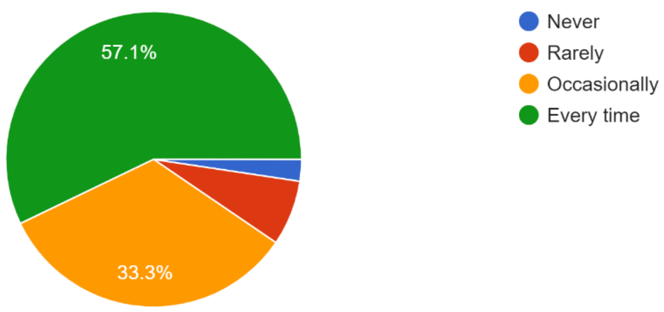
Findings



Participants usually review in all the ways possible when it comes to positive reviews. Negative reviews are mostly done using text and the star rating mechanism. Neutral reviews are mostly done by only text. However, there are occasions where users review with all possible methods in every aspect of opinions. There are also cases where users review with text, emojis, or a combination of both.

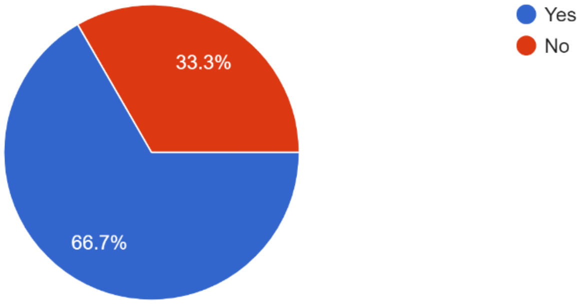
Question	**How often do you consider reading reviews before purchasing a product?**
Aim	To identify whether a user read other user's reviews about a certain product

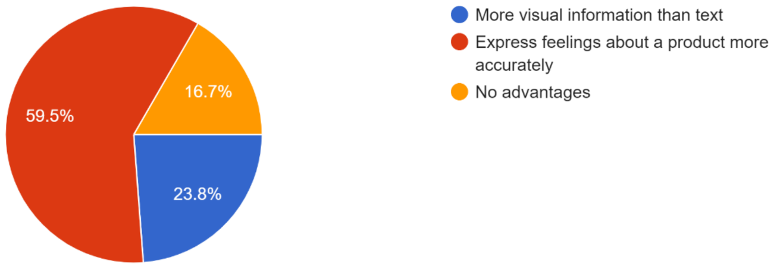
Findings

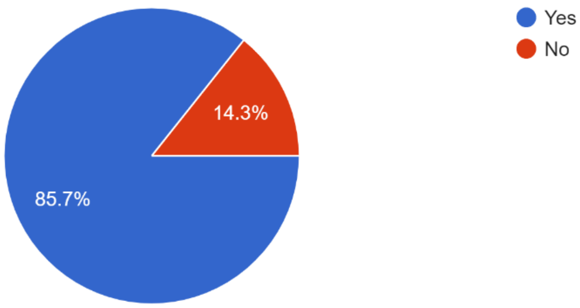


Most of the time, users read reviews before purchasing a product. There is also a possibility of not reading reviews before purchasing.

Question	**Have you ever encountered emojis in reviews?**
Aim	To identify whether a participant has ever come across emojis in a product review.

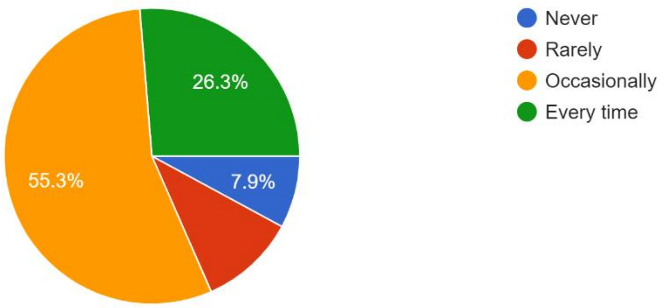
<table>
<tr><td colspan="2">Findings


According to the responses, there are occasions where the users has come across emojis in reviews. This concludes that there are reviews with emojis. There can be also emoji only reviews as well.</td></tr>
<tr><td>Question</td><td>What do you think are the advantages of having emojis in product reviews?</td></tr>
<tr><td>Aim</td><td>To identify the most valuable advantages of emojis in product reviews.</td></tr>
<tr><td colspan="2">Findings


According to the responds, reviews with emojis expresses feelings about a product more accurately while also adding more visual meaning than just using text.</td></tr>
<tr><td>Question</td><td>Do you think a recommendation system which analyzes emojis in reviews will be an improvement for product ranking?</td></tr>
<tr><td>Aim</td><td>To identify whether the proposed system will be useful in it's primary process.</td></tr>
<tr><td colspan="2">Findings


Majority of the responders aggress that the proposed system is useful for online platforms.</td></tr>
<tr><td>Question</td><td>If yes, how often would you use the proposed solution while deciding to purchase a product?</td></tr>
<tr><td>Aim</td><td>To identify the usage frequency of the proposed system.</td></tr>
</table>

Findings



It can be concluded that this system will be used occasionally by the users. There is also a possibility of using this system every time when users purchasing a product.

Question	**Any suggestions to improve this system?**
Aim	To identify additional requirements or improvements for the system

Findings

Some responders have suggested to analyze voice feedbacks with emoji interpretation. Participants also suggests that filtering products based on emojis according to the current emotion will also be an improvement.

Table 8: Survey findings

4.6. Summary Of Findings

Findings	**Survey**	**Literature Review**
The sentiment value of emphasized words by repeating letters or intentional misspellings will be very important because currently all of them fall into the neutral sentiment opinion category. If they were properly preprocessed to get the actual words, it will affect and improve the sentiment values in a text.		x
Sentiment value of emojis need to be considered when it comes to product ranking through sentiment analysis of product reviews to improve the correctness of recommendation systems.		X
A system which considers explicit and dynamic product aspect level sentiments would be able to analyze the sentiment score more accurately.		X
Emoji analysis can improve the overall sentiment analysis process.	X	X
Emojis provide better visual information and they are the preferred way of expressing feelings about a product.	X	
Sarcasm detection and negation handling must be considered when		X

developing a sentiment analysis-based system.		
External influence can be taken into consideration when generating recommendations.		X

Table 9: Summary of findings

4.7. Context Diagram

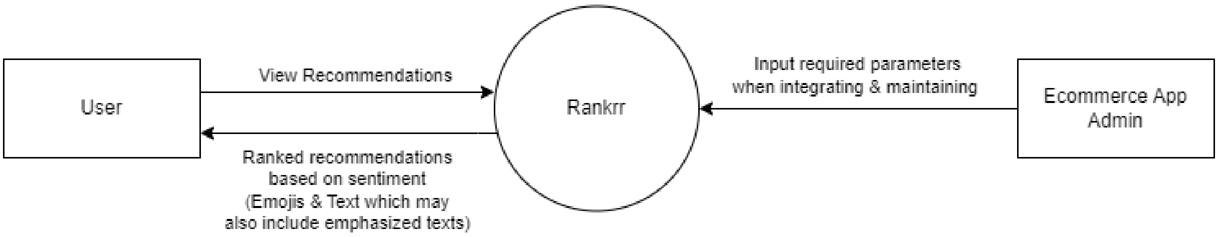


Figure 5: Context diagram (Level 0 Data Flow)

4.8. Use Case Diagram

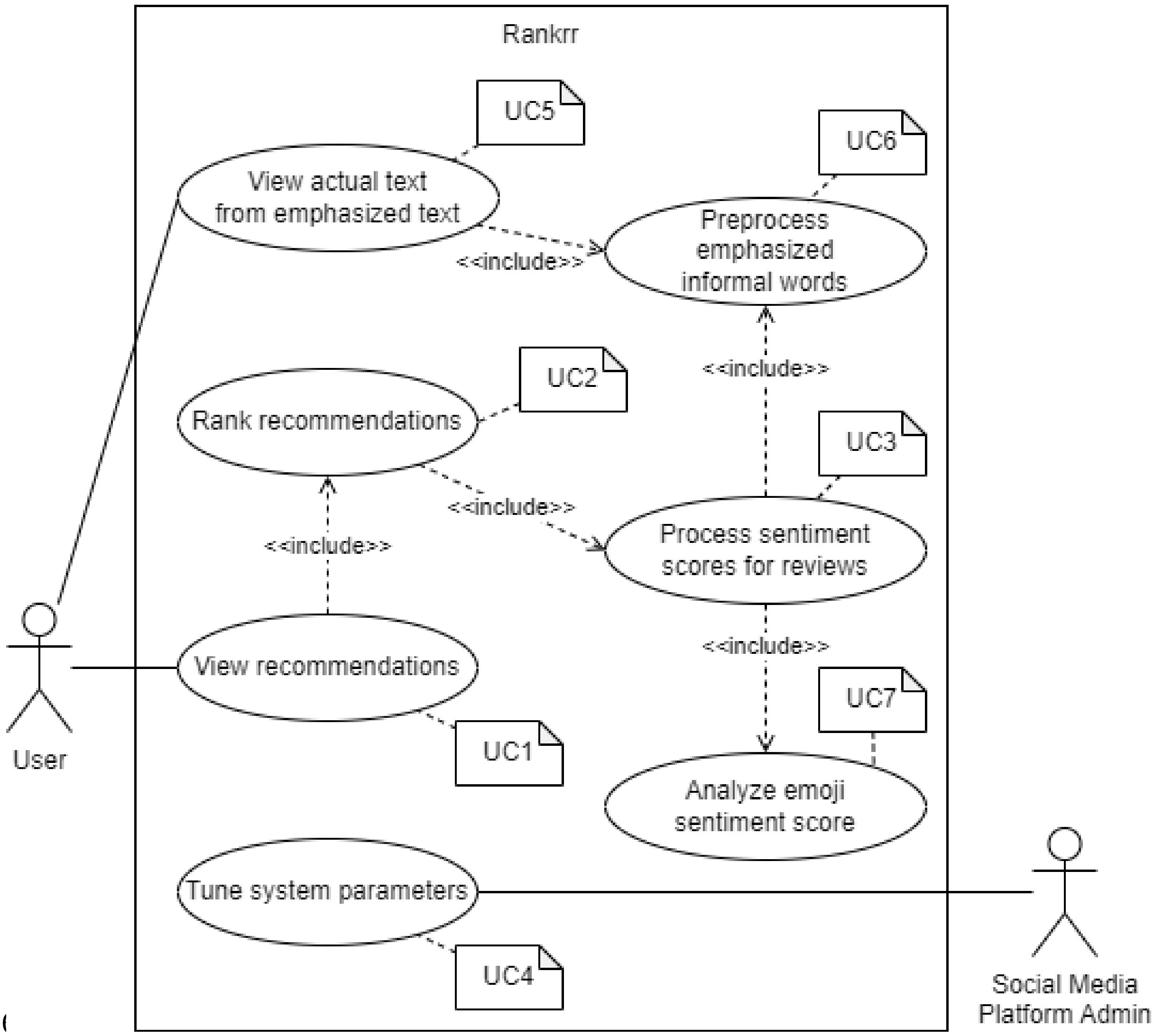


Figure 6: Use Case diagram

4.9. Use Case Descriptions

Use Case	Rank recommendations
ID	UC2
Description	Rank the given recommendations based on sentiment of text and emojis in product reviews.
Pre-Conditions	Should have current recommendations and the product reviews data.
Included use cases	UC3
Trigger	When a user view recommendations.
Main Flow	1. Sort recommendations based on sentiment score. 2. Provide the sorted list.
Exception Flow	Provide recommendations as is without any custom sorting.

Table 10: Use case description for UC2

Use Case	Process sentiment scores for reviews
ID	UC3
Description	Process sentiment scores for reviews based on text, emojis and emphasized text content.
Included use cases	UC6, UC7
Trigger	When ranking process has been triggered.
Main Flow	1. Preprocess each review text to clean and remove noise. 2. Calculate the positive sentiment scores for each review. 3. Calculate the mean sentiment score for each product in the given recommendations list through the scores from step 2.
Exception Flow	Don’t calculate sentiment scores.

Table 11: Use case description for UC3

Use Case	Preprocess emphasized informal words
ID	UC6
Description	Preprocess emphasized words to get the actual words which are meant by them.
Trigger	When ranking sentiment analysis process has been triggered through UC2 or UC5 which is a separate emphasized text preprocessing and viewing action.
Main Flow	1. Identify words with repeating letters in the given text. 2. Preprocess identified words to get the actual words.

	3. Replace the identified words in the given text with actual words which were acquired through step 2. 4. Return the cleaned text with actual words instead of words with repeated letters.
Exception Flow	Don’t clean the text and provide as it was given.

Table 12: Use case description for UC6

4.10. Requirements

Priority	**Description**
M - Must have	Primary functional requirements which must be implemented.
S - Should have	These requirements are not necessary for the primary functionality, but they are valuable to the system.
C - Could have	Desirable requirements which are optional and won’t have a critical effect if not implemented.
W - Will not have	Requirements which are not considered for the implementation, and these will not be implemented at the time.

Table 13: Requirement priority levels

4.10.1. Functional Requirements

ID	**Requirement**	**Priority**	**Use Case**
FR1	Users must be able to provide the required product recommendations with the reviews so that they can view the ranked products based on sentiment scores of product reviews.	M	UC1
FR2	Must be able to rank a set of product recommendations based on the calculated sentiment scores of reviews.	M	UC2
FR3	The system should be able to calculate sentiment scores for product reviews which have emojis and text which may also contain emphasized words to get an overall sentiment score.	M	UC3
FR4	Preprocess emphasized words in a text content to get the text with actual words instead of emphasized words.	M	UC6
FR5	Admins should be able to modify system parameters to perform as needed.	C	UC4

Table 14: Functional requirements

4.10.2. Non-Functional Requirements

ID	Requirement	Description	Priority
NFR1	Performance	Should be able to perform well under heavy load on the system. Since the system is basically about array processing and sorting, it should be performant to provide results efficiently.	S
NFR2	Scalability	The system should be able to process concurrent ranking processes since multiple will be requesting the ranking mechanism at the same time.	C
NFR3	Extensibility	The system should be able to easily integrate new functionalities to improve the overall process. It should be also be able to easily integrate into a currently existing ranking system to improve that system's process.	S
NFR4	Quality	Quality of the ranking process and sentiment calculation process can be of maximum level possible utilizing relevant techniques and data.	C

Table 15: Non-Functional requirements

4.11. Chapter Summary

The requirement gathering has been documented in this chapter. Stakeholders were analyzed and relevant requirement gathering techniques was also documented. The findings of the survey and literature review process has been summarized. The primary functional requirements and non- functional requirements are also detailed with relevant diagrams.

5. Social, Legal, Ethical And Professional Issues

5.1. Chapter Overview

Social, legal, ethical, and professional issues which can happen in different phases of the research have been defines in this section along with the mitigation steps of them.

5.2. Issues And Mitigation

Issue	**Mitigation**
Social	• The survey questionnaire was created in the way that they do not require any personal information. • The consent of evaluators was taken to include their evaluation, name and designation through the questionnaire.
Legal	• All the used core technologies which include programming languages, libraries, and frameworks, are open-source tools. • Personal data will not be needed to perform operations in the system
Ethical	• There was no plagiarism or information fabrication in the documentation or the research. • All the information gained through existing studies was cited. • Participants of survey questionnaire were notified about this research project and their contribution for it.
Professional	• All the applications which were used to develop were either open source or community versions of proprietary software. • Student licensed applications were also used on several occasions. • The research project was documented thoroughly and attempted to make it a quality research according to research standards.

Table 16: SLEP Issues & mitigation

5.3. Chapter Summary

This section described about social, legal, ethical & professional issues which identified throughout the research.

6. DESIGN

6.1. Chapter Overview

This section will describe the design aspect of the system. The selected system architecture is explained with diagrams and the design decisions are justified in this section.

6.2. Design Goals

Design Goal	Description
Performance	Should be performant since large text data array processing is the primary computation task. Asynchronous and concurrency would be beneficial for the system in the performance aspect.
Scalability	System should be able to handle concurrent processing tasks since multiple users will be using the service.
Extensibility	Should be easy to extend the system by adding new functionalities as needed. The system also should be able to integrate easily with another system to extend that system.
Quality	The quality of the sentiment analysis and ranking process should be at maximum level possible to ensure users will get most relevant recommendations.

Table 17: Design goals

6.3. High Level Design

6.3.1. Architecture Diagram

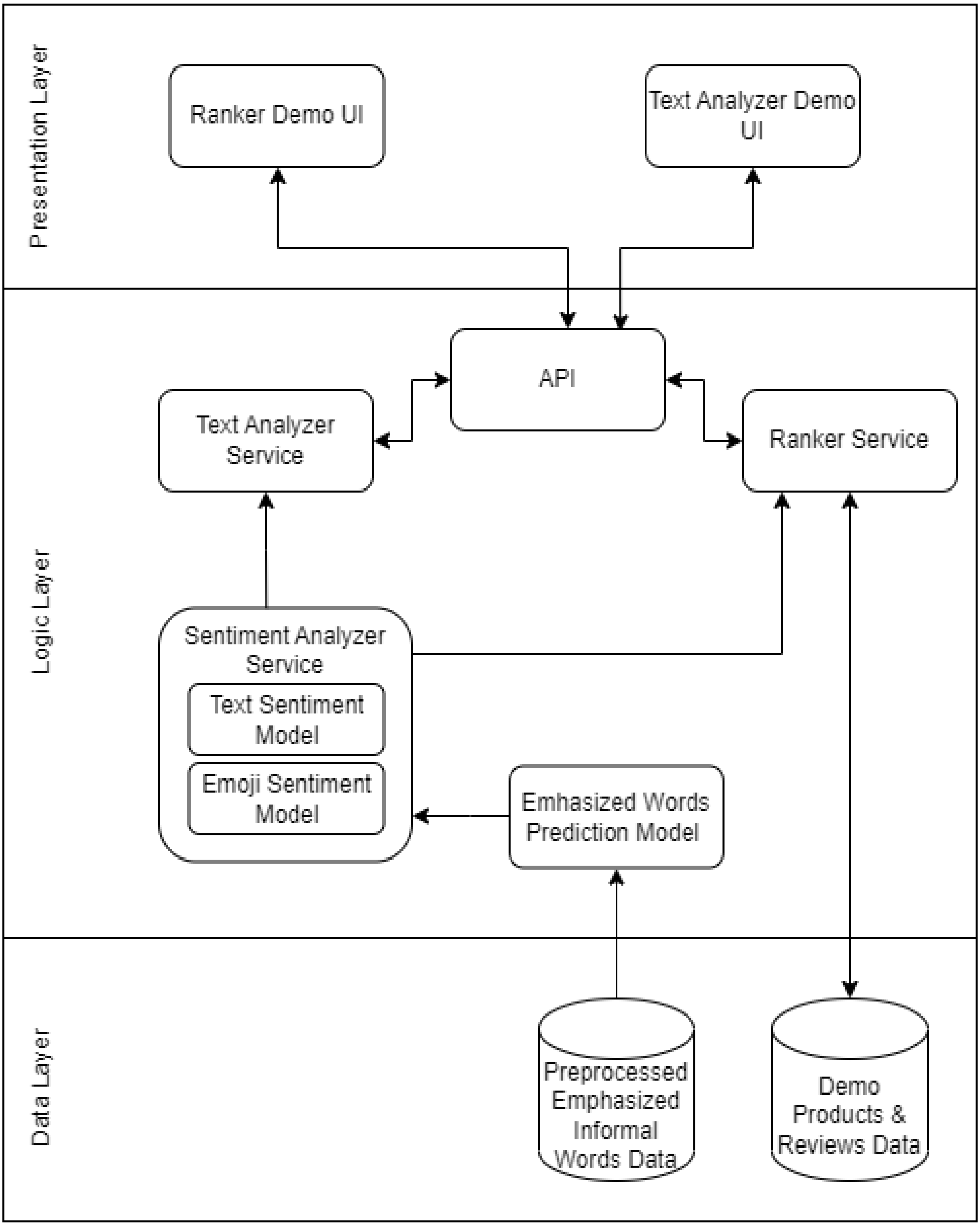


Figure 7: High level architecture diagram

6.3.2. Discussion Of Layers

Data Layer

- Emphasized Informal Words Data - The primary dataset which will be used to train the primary data science component which correctly predicts the actual word.
- Demo Products & Reviews Data - Used to analyze product reviews to identify emphasized words and to demonstrate the emoji & text sentiment based ranking methodology.

Logic Layer

- Emphasized Words Prediction Model - The component which will predict the actual word from a given emphasized informal word through text classification.
- Text Analyzer Service - Emphasized informal text preprocess functionalities resides here.
- Ranker Service - Contains primary functions which process, and rank a given set of products and reviews based on sentiment analysis.
- Sentiment Analyzer Service - Contains sentiment analysis models and functionalities.
- API - Primary entry point for the backend and will be used by frontend application.

Presentation Layer

- Text Analyzer Demo UI - A demo UI to showcase the possible existing methods and the approach of this research regarding emphasized informal text preprocessing.
- Ranker Demo UI - A UI to demonstrate the product recommendations ranking mechanism.

6.4. Low-Level Design

6.4.1. Choice Of Design Paradigm

SSADM (structured systems analysis and design methodology) has been chosen due to the following reasons:

- Primary component is a data science component which are more incline toward function-based approach.
- Easy to implement and extend as needed with more functionalities.

6.4.2. Design Diagrams

Level 1 Data Flow Diagram

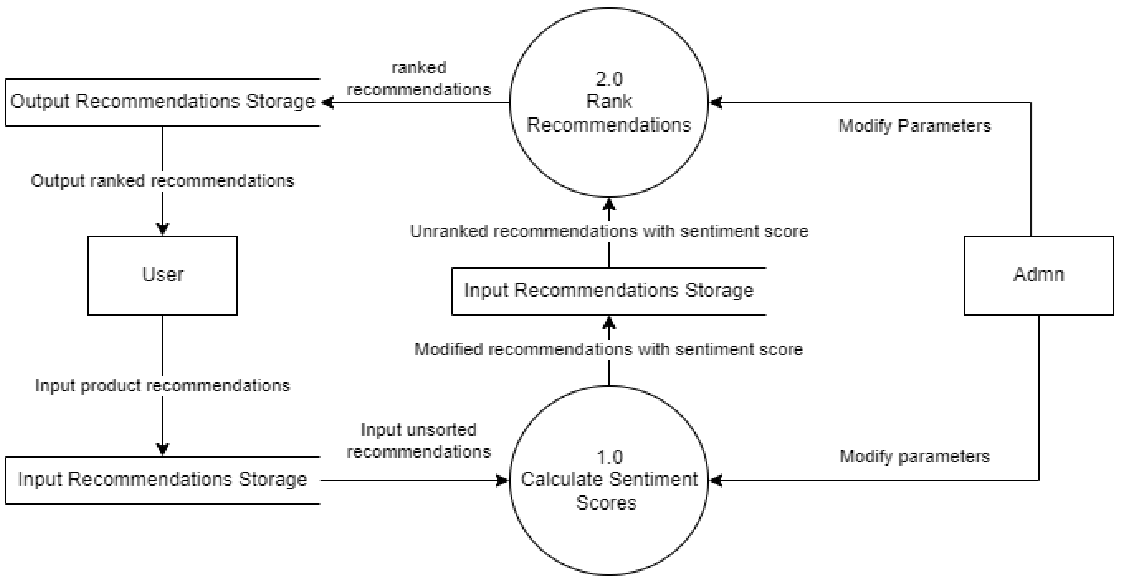


Figure 8: Level 1 Data Flow diagram

6.4.3. System Process Flow Chart

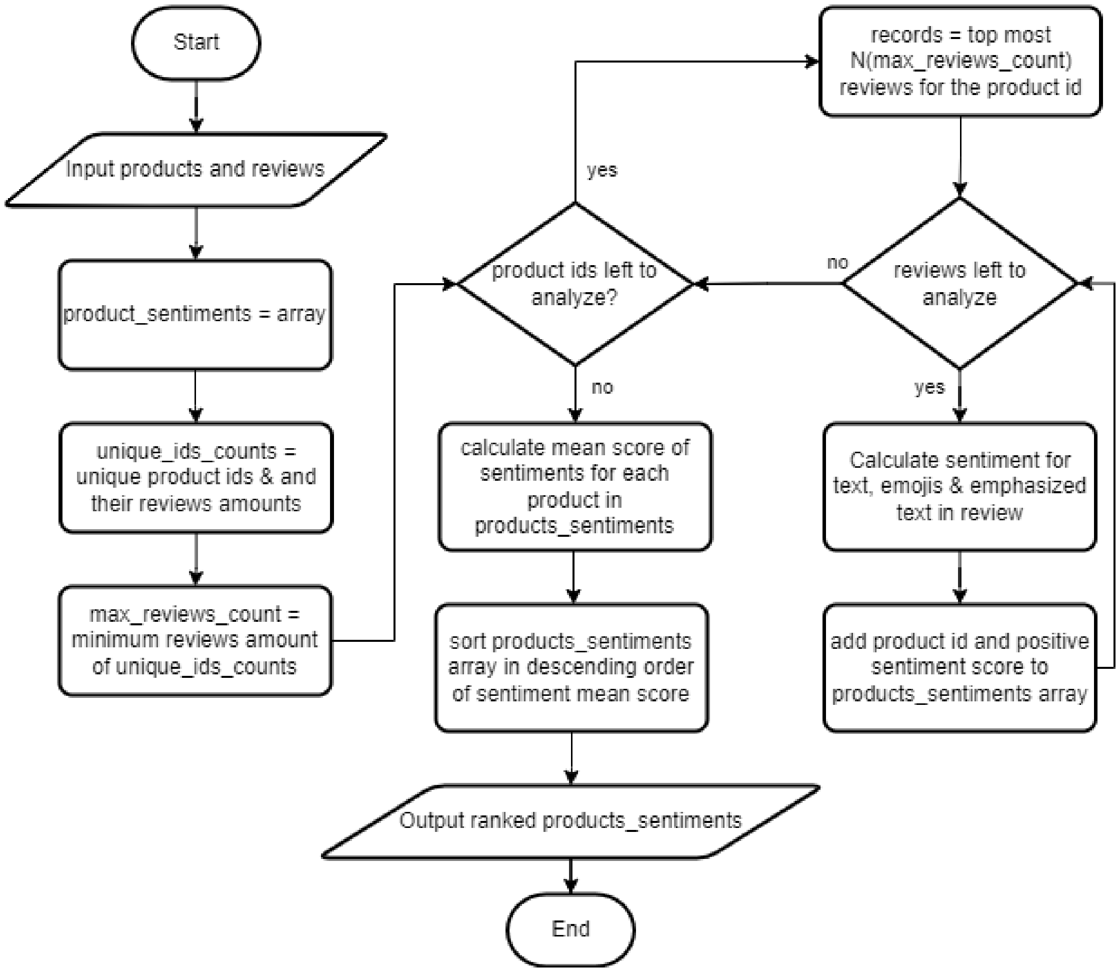


Figure 9: System process flow chart

6.5. Chapter Summary

This section summarized the primary design of the system. Architecture diagrams and primary process flow chart have been included.

7. IMPLEMENTATION

7.1. Chapter Overview

This chapter is about tools and various technologies which will be used to implement the system while also describing the implementation of the core functionality. The programming languages, libraries, frameworks, and other relevant tools were selected and justified.

7.2. Technology Selection

7.2.1. Technology Stack

Presentation Layer
HTML, TS
Logic Layer
Flask, scikit learn, NLTK
Data Layer
amazon, Synthesized Data of Emphaszed Text

Table 18: Technology stack

7.2.2. Dataset Selection

The following resources has been selected as the primary data sources to get the required user reviews on products and to analyze emphasized informal texts.

- Amazon customer reviews.
- Twitter Sentiment140 dataset.
- IMDB Movie Reviews.
- Emphasized texts dataset (Manually Synthesized Dataset).

7.2.3. Programming Languages

Language	Justification
Python	Selected as the primary programming language since the project is of data science field. Python has a set of most used libraries which has been utilized for data science tasks. It's also a very simple but powerful programming language when it comes to scientific computing. The data science component and it's API will be developed using Python.
TypeScript	Will be used to implement the required frontend application with angular framework. It will also be used to communicate with the Python based API.

Table 19: Programming languages

7.2.4. Libraries

Library	Justification
NLTK	For tokenization, POS tagging, lemmatization, and other related tasks for language processing.
Scikit-learn	Used for primary machine learning algorithms which will be needed to implement the required model.
Pandas	For dataset manipulation
Numpy	To work with arrays and matrices easily.
Matplotlib	Has visualizations for data science component analysis.

Table 20: Libraries

7.2.5. Development Frameworks

Framework	Justification
Flask	To build a minimal API based on Python
Bootstrap	For frontend CSS styling which will be used to style the UI
Angular	For frontend development of the demo web app.

Table 21: Frameworks

7.2.6. IDEs

IDE	Justification
PyCharm	To develop and test the data science component. Easy to generate required visualizations.
VSCode	To develop the Python & Flask based API and the Frontend. Has a set of extensions which will be useful in the process. A very simple and resource friendly application.

Table 22: IDEs

7.2.7. Summary Of Technology Selection

Component	Tools
Programming Languages	Python, TypeScript
Backend Frameworks	Flask
Frontend Frameworks	Angular, Bootstrap
Libraries	NLTK, Scikit-learn, Pandas, Numpy, Matplotlib
IDE	PyCharm, VSCode
Version Control	Git, GitHub

Table 23: Technology selection summary

7.3. Core Functionality Implementation

The goal of the implementation is to develop a sentiment analysis based products ranking solution to improve the ranking of a given set of recommendations by considering sentiment of both the emojis and text in user reviews which may also include emphasized informal text with repeating characters or misspellings while also being able to preprocess these kinds of text through a machine learning based word prediction model.

7.3.1. Emphasized Informal Text Preprocessing With Machine Learning

Words such as “liiiike”, “puuurfect” or “uhhhhhhh” are commonly occurs in user reviews on ecommerce platforms. The preprocessing task of these words is to correctly get the actual words which are meant by them (such as “like”, “perfect”, “ugh”). However, to create a machine learning based solution, a dataset is needed. Since there weren’t any publicly available datasets, the synthetization of a custom dataset is mandatory.

Figure 10: Word cloud for emphasized words

To do so, the Amazon reviews, IMDB movie reviews, and Twitter sentiments datasets were analyzed to identify these kinds of words and display a word cloud to easily investigate them.

Several other analyzed datasets are included in the **Appendix - 1**

Synthesizing Dataset

```
# words = ["like", "beautiful", "love", "cool", "waste", "boring", etc....]
dataset = []
for word in words:
    for num in range(1000000):
        new_word = ""
        repeat_letter_index = random.randint(0, len(word)-1)
        repeat_letter = word[repeat_letter_index]
        repeat_count = random.randint(2, 15)
        for x in range(0, len(word)):
            if x == repeat_letter_index:
                new_word += repeat_letter * repeat_count
            else:
                new_word += word[x]
        dataset.append({"emphasized": new_word, "actual": word})

df = pd.DataFrame.from_dict(dataset)
df = df.drop_duplicates(subset=['emphasized'])
df.to_csv("repeated_letters_words_v9.csv")
```

Figure 11: Code snippet - dataset synthetization

After the analysis, the actual words were collected manually and from that a synthetic dataset has been generated by randomly repeating one of the letters in a word. This process runs for a random amount of turns to all the collected words. These words are then put into an array of dictionaries which contains the generated word and its actual counterpart word. After that, the duplications are removed and saved to a csv file.

Text Classification Model

After that task, a classification algorithm based word prediction model has been developed. To get the best results, the dataset has been trained on several classification algorithms through a pipeline and selected the most performant one which was the Random Forest Classification based model.

```
df = pd.read_csv('repeated_letters_words_v9.csv')
df = df.sample(frac = 1)

pipeline = Pipeline([
    ('vect', CountVectorizer(analyzer='char')),
    ('tfidf', TfidfTransformer()),
    ('clf', MultinomialNB())
])
```

Figure 12: Code snippet - dataset load & pipeline define

With the use of scikit-learn's pipeline module, a model pipeline has been defined to train the actual model based on the dataset. First step is about tokenizing the emphasized words into each character with the use of 'CountVectarizer' with 'char' analysis mode. With that, the model can learn various patterns on the occurrences of different letters in the emphasized words. 'TfidfTransformer' has been used make the model focus on unique features of emphasized words. 'MultinomialNB' is the initial classifier to train the model.

```
parameters = {
    'tfidf__use_idf': [(True, False), (False, True), (True, True), (False, False)],
    'clf': [MultinomialNB(), DecisionTreeClassifier(), RandomForestClassifier(), SVC(), KNeighborsClassifier()],
}

X_train, X_test, y_train, y_test = train_test_split(df['emphasized'], df['actual'], test_size=0.2, random_state=42)

grid_search = GridSearchCV(pipeline, parameters, cv=5, n_jobs=-1)
grid_search.fit(X_train, y_train)

best_model = grid_search.best_estimator_
best_params = grid_search.best_params_
```

Figure 13: Code snippet - parameters, dataset split & estimator search

Then the 'GridSearchCV' is used to search over the given parameters and classifiers to get the best possible model by training on the dataset. The parameter 'tfidf__use_idf' is about applying inverse document frequency weighting to the letter frequencies and L2 normalization for importance of each letter for an emphasized word. All the combinations of applying these two techniques has been defined, so that the best combination can be identified and used. The parameter 'clf' is about the classifier algorithm which in this case is a set of multi class classifiers. With that, the best possible classifier can be identified through the 'GridSearchCV'.

7.3.2. Sentiment Based Ranking With Emoji & Emphasized Informal Text

The ranking process is about sorting a set of products based on their sentiments of reviews while considering the emojis and emphasized informal words in these reviews since they are valuable to the sentiment analysis task.

```
counts_dict = df[product_id_col].value_counts().to_dict()
uniques_and_counts = dict(sorted(counts_dict.items(), key=lambda x: x[1], reverse=True))
uniques_and_counts = dict(list(uniques_and_counts.items())[:max_products_amount])
max_reviews_amount = min(uniques_and_counts.values())
```

Figure 14: Code snippet - reviews filtering

The first task is to get the unique product ids and the amount of their occurrences from the given data while also considering a maximum number of products which will be displayed in the end after the ranking. To make the system fair, a maximum number of reviews per product id has been calculated by getting the minimum occurrences amount from the previous list of items which contains product ids, and their occurrences amounts.

```
sentiments = {}
if consider_emoji:
    sia = emoji_sia
else:
    sia = text_sia

for product_id in uniques_and_counts.keys():
    records = get_top_records(df, product_id, product_id_col, max_reviews_amount)
    for i, record in records.iterrows():
        record[review_text_col] = record[review_text_col].lower()
        if consider_emoji:
            emojis = emoji_util.extract_emojis(record[review_text_col])
        record[review_text_col] = emoji_util.remove_emojis(record[review_text_col])
        if consider_emph_text:
            record[review_text_col] = text_analyzer.get_ml_preprocessed_text(record[review_text_col])
        record[review_text_col] = sentiment_analyzer.preprocess_text(record[review_text_col])
        if consider_emoji:
            record[review_text_col] = record[review_text_col] + " " + " ".join(emojis)
        if product_id in sentiments:
            sentiments[product_id].append(sia.polarity_scores(record[review_text_col])['pos'])
        else:
            sentiments[product_id] = [sia.polarity_scores(record[review_text_col])['pos']]
```

Figure 15: Code snippet - calculating sentiments for reviews on each product

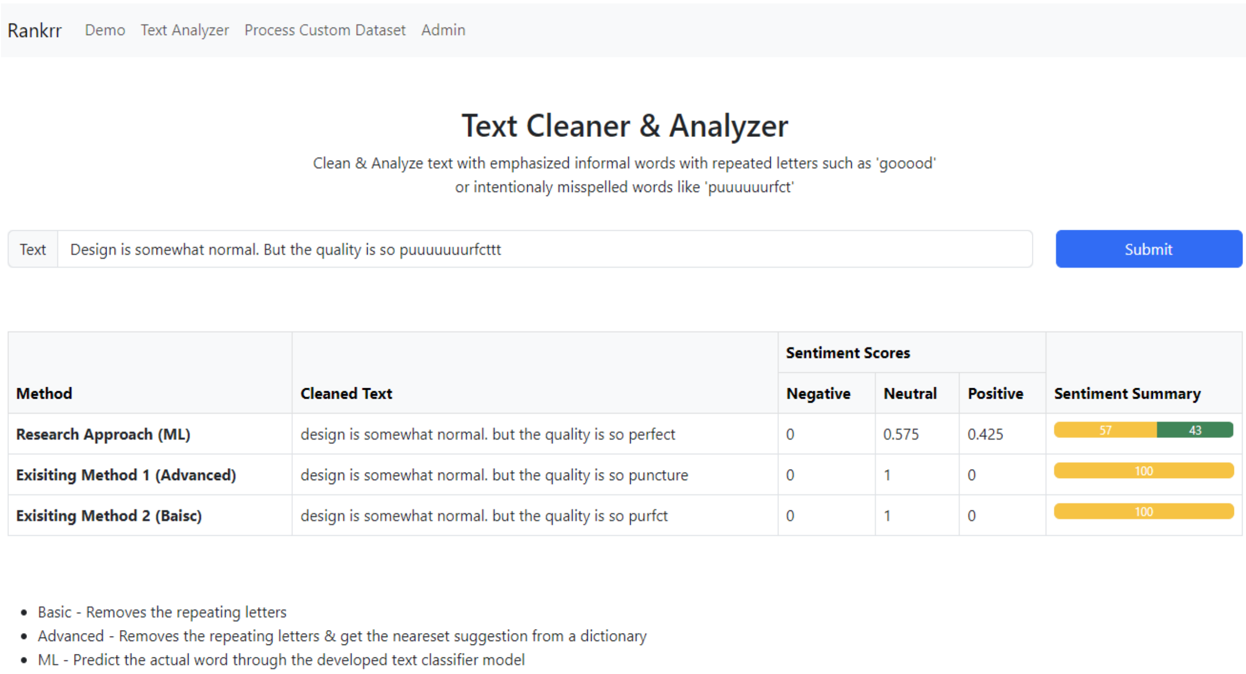
After that, for each of the product ids which was gotten from the previous step, the topmost records (while respecting the maximum number of reviews) will be extracted from the given data. For each of the reviews from these records, a set of preprocessing steps will occur. At first, the text will be converted to lower case. If the condition to consider emojis is true, the emojis will be separated from the text and the other preprocessing steps will follow. If the condition to preprocess the emphasized words is true, the developed text classifier model will be used to correctly clean these words. Then, the standard preprocessing tasks such as stop words removal and lemmatization will happen to prepare the text for the sentiment analysis models. The previously separated emojis will be joined at the end of the review text. If the emoji consideration condition is true, the emoji and text sentiment model will be used to get the positive sentiment score for the review text. In the case of this condition is false, the text only sentiment model will be used to get the positive sentiment score. Anyhow, the currently iterating product id will be stored as a key in a dictionary and the review sentiment score will be appended to an array which will be the value in the dictionary. This will iteratively happen for all the relevant reviews and product ids.

```
for key, value in sentiments.items():
    sentiments[key] = np.mean(value)
sorted_data = sorted(sentiments.items(), key=lambda x: x[1], reverse=True)
return [{'product_id': key, 'sentiment_score': val, 'rank': rank} for rank, (key, val) in enumerate(sorted_data, start=1)]
```

Figure 16: Code snippet - Ranking products on mean sentiment score

After that, the mean sentiment score will be calculated from the array of sentiment scores for each of the product ids and will be sorted in descending order. Using the list comprehension method, a list of dictionaries will be created in which each dictionary contains the product id, sentiment score, and the rank (starts from 1). This list will be returned since it is the ranked set of products based on sentiment analysis of emojis and emphasized text interpretation.

7.4. User Interface



Method	Cleaned Text	Sentiment Scores			Sentiment Summary
		Negative	Neutral	Positive	
Research Approach (ML)	design is somewhat normal. but the quality is so perfect	0	0.575	0.425	57 / 43
Exisiting Method 1 (Advanced)	design is somewhat normal. but the quality is so puncture	0	1	0	100
Exisiting Method 2 (Baisc)	design is somewhat normal. but the quality is so purfct	0	1	0	100

- Basic - Removes the repeating letters
- Advanced - Removes the repeating letters & get the neareset suggestion from a dictionary
- ML - Predict the actual word through the developed text classifier model

Figure 17: Emphasized text preprocessor & analyzer demo

Rankrr Demo Text Analyzer Process Custom Dataset Admin

Rankrr Demo

Compare the ranking differences of two methodologies. Click on a product ID in first table to compare the reviews sentiment analysis with the two methodologies

With Emoji & Emphasized Text Interpretation

Rank	Product ID	Sentiment Score	Rank Change
1	B00F9RJLTU	0.43162500000000004	0
2	B00GBT8QM8	0.3701875	0
3	B00JTYYLJO	0.3696875	1
4	B00KJIVDW2	0.3644375	3
5	B00ECIYWA0	0.3628125	-2
6	B00ESMSCWO	0.34431249999999997	6
7	B00L8WBEMW	0.3299375	-2
8	B00CONNILE	0.31912500000000005	2
9	B00JUW0IM4	0.29225	-1
10	B00LLIVQM6	0.2848125	-1
11	B00K2QLZMO	0.2820625	0
12	B00CONNDGO	0.27656250000000004	-6
13	B00CONNGVQ	0.23712500000000003	0
14	B000NZW3IY	0.2115	0
15	B002HJ377A	0.1496875	0

Without Emoji & Emphasized Text Interpretation

Rank	Product ID	Sentiment Score
1	B00F9RJLTU	0.41218750000000004
2	B00GBT8QM8	0.39325
3	B00ECIYWA0	0.39306250000000004
4	B00JTYYLJO	0.3815625
5	B00L8WBEMW	0.37912500000000005
6	B00CONNDGO	0.32781249999999995
7	B00KJIVDW2	0.31725000000000003
8	B00JUW0IM4	0.30393749999999997
9	B00LLIVQM6	0.2829375
10	B00CONNILE	0.2814375
11	B00K2QLZMO	0.27431249999999996
12	B00ESMSCWO	0.2495
13	B00CONNGVQ	0.19637500000000002
14	B000NZW3IY	0.189375
15	B002HJ377A	0.155125

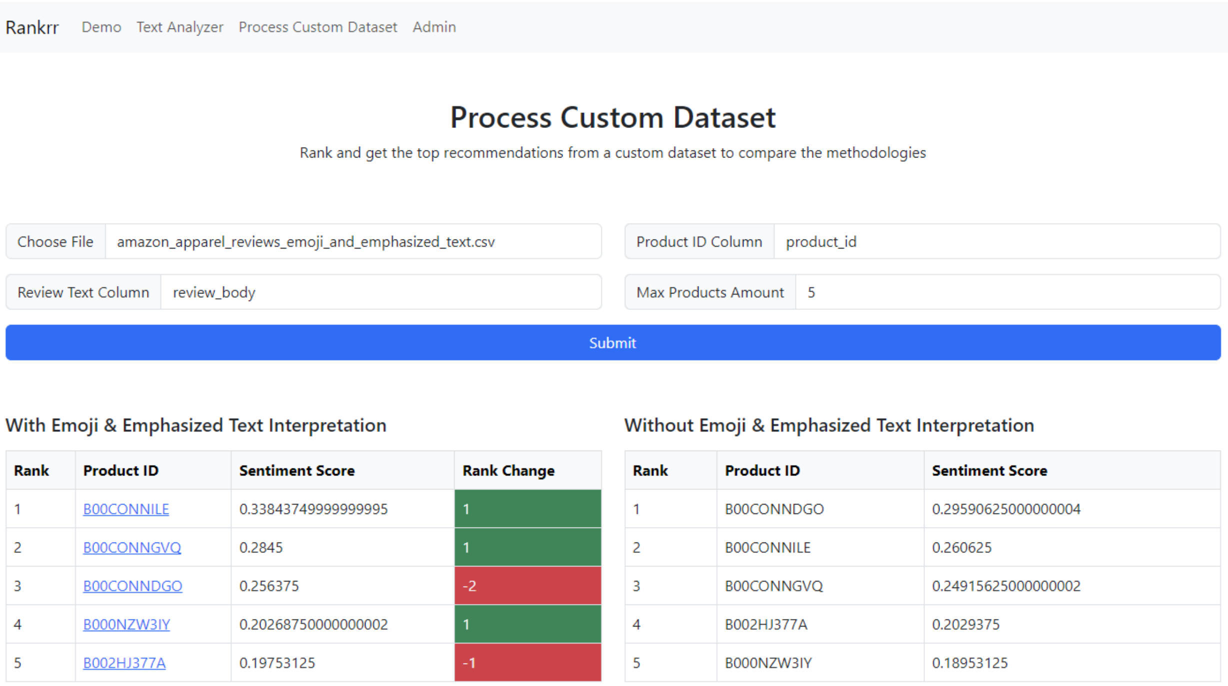
Figure 18: Sentiment based ranker demo

Reviews Sentiments - B00ESMSCWO

Review	With Emoji & Emphasized Text Interpration				Without Emoji & Emphasized Text Interpration			
	Sentiment Scores			Sentiment Summary	Sentiment Scores			Sentiment Summary
	Negative	Neutral	Positive		Negative	Neutral	Positive	
This runs wayyy to small , go by your waist inches not the size of your clothes. But great product I love it, and I love the fast service , I did returned it and order a diferent size !	0	0.643	0.357	64 36	0	0.701	0.299	70 30
I love it😍 im 5'2 and i weight 110lbs. I get the small n its perfect. My waist is 28. Theres a lil fat hanging below my bra but thats ok it looks good on me.	0	0.6	0.4	60 40	0	0.699	0.301	70 30
Okay, I read so many of these reviews that my head was spinning. I bought a size 38 and it is WAY wayyyy too small. I am 5'3", about 159 lbs, (but carry most of my weight around my middle) and a size 10 in pants. The 38 wasn't going on, even with my husbands help. It is a good quality item, though. I'm reordering another one.	0	0.874	0.126	87 13	0	0.885	0.115	89 12
I loveee😍😍😍	0	0.405	0.595	41 60	0	1	0	100
I love this waist trainer although it runs toooo small, I am a size small therefore I ordered a small size but ill say the size small is more like an xxs.	0.106	0.642	0.253	11 64 25	0.109	0.706	0.185	11 71 19
😄	0	0.353	0.647	35 65	0	0	0	
Ummmmmmmm definitely measure yourself before ordering these... I ordered this one and I couldn't even come close to closing it LOL That's my fault though. I ended up ordering another one in a larger size that I love and when I get smaller, I'll switch to this one. :) I love these!!	0.052	0.601	0.347	5 60 35	0.046	0.659	0.295	5 66 30
I was so excited to get this product, when I got it, it was sooo small. I'm 5'2 and weigh 115 lbs, I ordered a size small and it did not fit at all. I am returning it and ordering a new one!	0	0.822	0.178	82 18	0.053	0.873	0.074	5 87 7
I'm 5'2" 135 lbs and I got a small size and it fits perfect. It's a little snug but that's the point of it. I can't wait to get to the next hook. This helps me control my appetite so that I don't over eat. Also this makes my back feel amazing 😍	0	0.753	0.247	75 25	0	0.792	0.208	79 21
👍	0	1	0	100	0	0	0	
But i love it 😍 thank you	0	0.236	0.764	24 76	0	0.249	0.751	25 75
Waaaaaaay to small!! I even went up 2 sizes don't waste your money	0	0.774	0.226	77 23	0	0.774	0.226	77 23

Close

Figure 19: Sentiment on reviews demo



Rank	Product ID	Sentiment Score	Rank Change
1	B00CONNILE	0.33843749999999995	1
2	B00CONNGVQ	0.2845	1
3	B00CONNDGO	0.256375	-2
4	B000NZW3IY	0.20268750000000002	1
5	B002HJ377A	0.19753125	-1

Rank	Product ID	Sentiment Score
1	B00CONNDGO	0.29590625000000004
2	B00CONNILE	0.260625
3	B00CONNGVQ	0.24915625000000002
4	B002HJ377A	0.2029375
5	B000NZW3IY	0.18953125

Figure 20: Custom dataset ranking demo

7.5. Chapter Summary

This section explained the tools and technologies which are used to implement the system and justifications has been provided for the selection of them. It also described the core functionality implementation and the user interface of the demo application.

8. TESTING

8.1. Chapter Overview

This section covers the overall testing process of the research. It also describes about the core data science component testing, evaluation, and benchmarking.

8.2. Objectives And Goals Of Testing

- To make sure that the data science model functions as intended
- Most important functional requirements are met by the system.
- Benchmark the system to compare with other systems to improve upon it.
- Find and prevent defects to maintain an optimum system.
- Provide stakeholders with testing information to make decisions regarding the system.

8.3. Testing Criteria

- Functional - Regarding the design specifics and the actual implementation of the system in regard to specified functional requirements.
- Non-Functional - Non-functional aspects of the system such as performance, scalability, and extensibility.

8.4. Model Evaluation

The emphasized informal text classification model has been evaluated with standard evaluation metrics.

8.4.1. Classifier Results

```
Best Model: Pipeline(steps=[('vect', CountVectorizer(analyzer='char')),
                ('tfidf', TfidfTransformer(use_idf=(True, False))),
                ('clf', RandomForestClassifier())])
Best Parameters: {'clf': RandomForestClassifier(), 'tfidf__use_idf': (True, False)}
```

Figure 21: Pipeline results

8.4.2. Classification Report

Accuracy Score: 0.9887459807073955

	Precision	Recall	F1-Score	Dataset Support
Macro Avg	0.98	0.98	0.98	622
Weighted Avg	0.99	0.99	0.99	622

Table 24: Standard evaluation metrics

8.4.3. Confusion Matrix

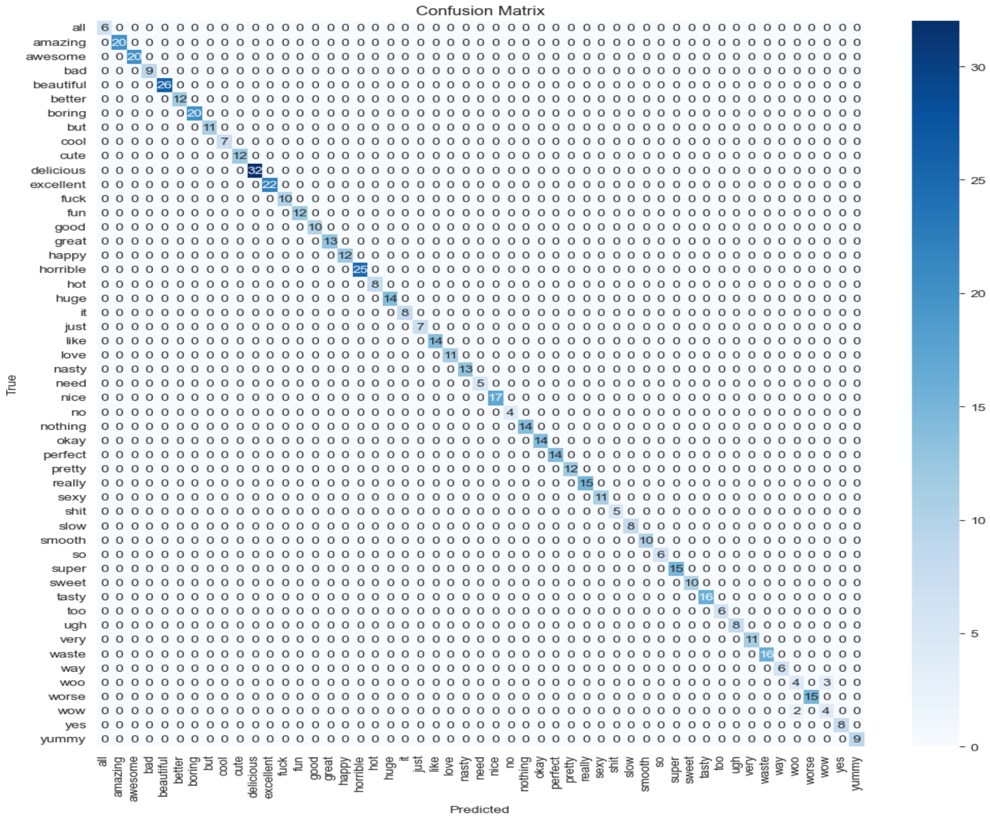


Figure 22: Confusion matrix

8.4.4. AUC / ROC Curve

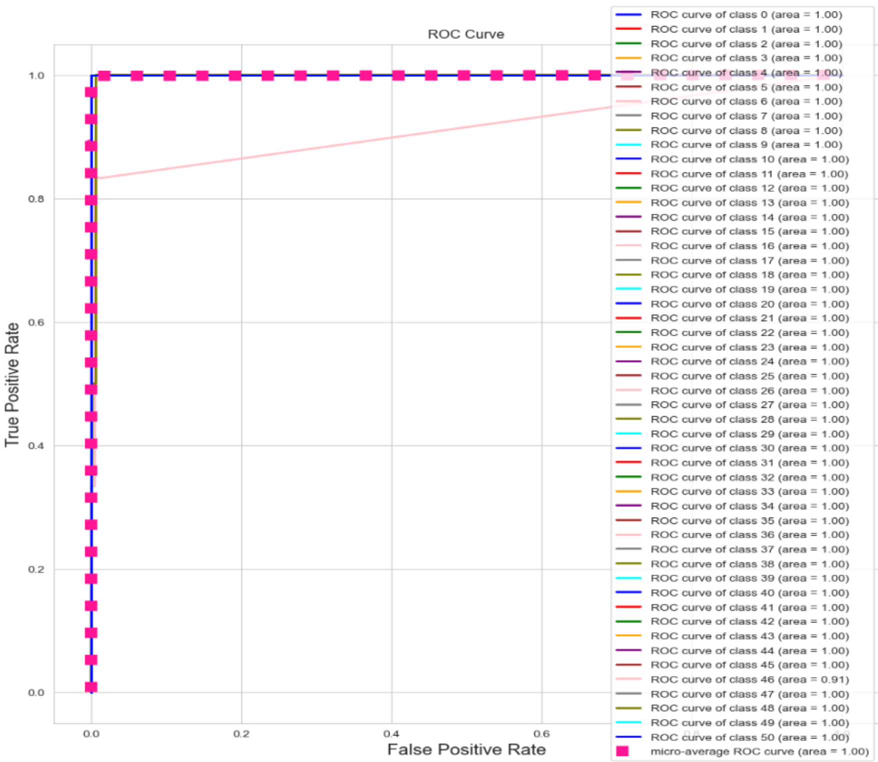


Figure 23: AUC / ROC Curve

8.5. Benchmarking

Text | Design is somewhat normal. But the quality is so puuuuuuurfcttt | Submit

Method	Cleaned Text	Sentiment Scores			Sentiment Summary
		Negative	Neutral	Positive	
Research Approach (ML)	design is somewhat normal. but the quality is so perfect	0	0.575	0.425	57 43
Exisiting Method 1 (Advanced)	design is somewhat normal. but the quality is so puncture	0	1	0	100
Exisiting Method 2 (Baisc)	design is somewhat normal. but the quality is so purfct	0	1	0	100

- Basic - Removes the repeating letters
- Advanced - Removes the repeating letters & get the neareset suggestion from a dictionary
- ML - Predict the actual word through the developed text classifier model

Figure 24: Benchmarking of emphasized text preprocessing

Benchmarking has been done by simulating existing non-data science methods with the proposed method to compare the emphasized words prediction model.

With Emoji & Emphasized Text Interpretation

Rank	Product ID	Sentiment Score	Rank Change
1	B00F9RJLTU	0.43162500000000004	0
2	B00GBT8QM8	0.3701875	0
3	B00JTYYLJO	0.3696875	1
4	B00KJIVDW2	0.3644375	3
5	B00ECIYWA0	0.3628125	-2
6	B00ESMSCWO	0.34431249999999997	6
7	B00L8WBEMW	0.3299375	-2
8	B00CONNILE	0.31912500000000005	2
9	B00JUW0IM4	0.29225	-1
10	B00LLIVQM6	0.2848125	-1
11	B00K2QLZMO	0.2820625	0
12	B00CONNDGO	0.27656250000000004	-6
13	B00CONNGVQ	0.23712500000000003	0
14	B000NZW3IY	0.2115	0
15	B002HJ377A	0.1496875	0

Without Emoji & Emphasized Text Interpretation

Rank	Product ID	Sentiment Score
1	B00F9RJLTU	0.41218750000000004
2	B00GBT8QM8	0.39325
3	B00ECIYWA0	0.39306250000000004
4	B00JTYYLJO	0.3815625
5	B00L8WBEMW	0.37912500000000005
6	B00CONNDGO	0.32781249999999995
7	B00KJIVDW2	0.31725000000000003
8	B00JUW0IM4	0.30393749999999997
9	B00LLIVQM6	0.2829375
10	B00CONNILE	0.2814375
11	B00K2QLZMO	0.27431249999999996
12	B00ESMSCWO	0.2495
13	B00CONNGVQ	0.19637500000000002
14	B000NZW3IY	0.189375
15	B002HJ377A	0.155125

Figure 25: Benchmark of sentiment based ranking

Also, the ranking process has been compared with a simulation of the differences of the proposed method and existing techniques.

8.6. Functional Testing

Test Case	FR ID	Trigger	Expected Result	Actual Result	Result Status
1	FR1	User provides a set of products and their reviews.	Ranked list of products based on the sentiment scores of reviews.	Ranked list of products based on the sentiment scores of reviews.	Passed
2	FR2	After user provides a set of products with reviews.	Rank the list of products based on the mean positive sentiment score of all reviews on each product.	Rank the list of products based on the mean positive sentiment score of all reviews on each product.	Passed
3	FR3	Ranking the products through sentiment scores.	Sentiment scores for each review on each product.	Sentiment scores for each review on each product.	Passed
4	FR4	When text with emphasized words needs to be preprocessed.	Preprocessed and cleaned text which replaces emphasized words that have repeating letters or misspelling with the actual words which are meant by them.	Preprocessed and cleaned text which replaces emphasized words that have repeating letters or misspelling with the actual words which are meant by them.	Passed
5	FR5	Change options to rank the products.	Adjusted ranking of the products with changed options.	Not Implemented	Failed

Table 25: Functional Testing

8.7. Module And Integration Testing

Module	Input	Expected Output	Actual Output	Result
Emphasized informal words preprocessor	Emphasized words with repeating letters and/or misspelled letters	Actual meaningful words which are meant by the informal words	Actual meaningful words which are meant by the informal words	Passed
Emoji and Text Sentiment Analyzer	Text content which are mixed with emojis	A valid sentiment score which also considers the emojis.	A valid sentiment score which also considers the emojis.	Passed
Products Ranker	Products and their Reviews	Ranked products list based on the positive sentiment of the overall reviews of each product.	Ranked products list based on the positive sentiment of the overall reviews of each product.	Passed

Table 26: Module & Integration Testing

8.8. Non-Functional Testing

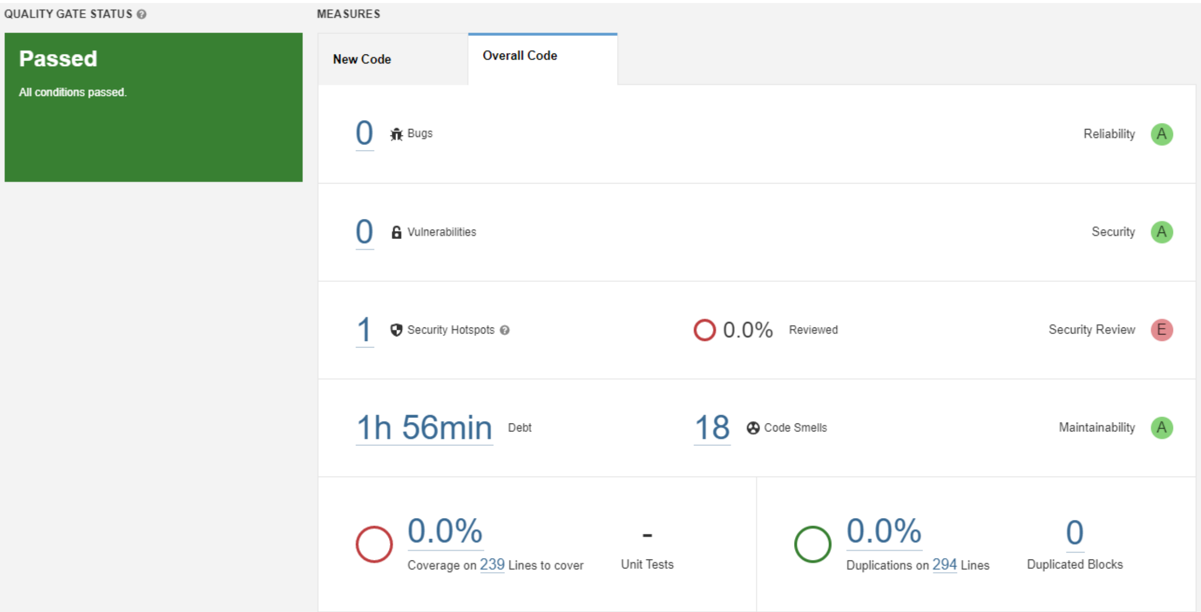


Figure 26: Non-Functional Testing - SonarQube scan evaluation on serverside code base

8.9. Limitations Of Testing

The primary limitation of testing was to benchmark the emphasized informal words prediction model since there were no existing methods which uses data science to solve this problem. Hence, the simulation of existing techniques which doesn't use data science has been compared with the proposed solution. Likewise, the ranking process has been compared with a simulation of the current methodology to showcase the differences and improvements.

8.10. Chapter Summary

This section described the testing, evaluation, and benchmarking of the main data science component and several functional and non-functional testing of the overall system.

9. EVALUATION

9.1. Chapter Overview

This chapter will discuss the overall evaluation of the implemented system. This chapter will summarize the evaluation process with its results and limitations.

9.2. Evaluation Criteria

Criteria	**Purpose**
Research choice	Assess the choice of research domains and the topics related to this research
Research **c**ontribution	Evaluate the contributions for the fields of recommendation systems and NLP domains.
Development	Validate the development process and the overall implementation
Improvements	Identify current limitations which can be addressed through future improvements.

Table 27: Evaluation criteria

9.3. Self-Evaluation

Criteria	Evaluation Summary
Research choice	The primary research domains, which are recommendation systems and NLP (Natural Language Processing) are two of the mostly researched domains. Also, NLP is a core component for development of artificial intelligence as well.
Research Contribution	There is a technical contribution for the domain of NLP since current studies related to preprocessing emphasized informal words, were not found so far. Using machine learning to solve this issue based on a synthesized dataset is a major contribution to the domain of NLP because these emphasized words hold valuable information specially for tasks such as sentiment analysis. The domain contribution is that this research considers emojis as well as emphasized informal text in product reviews so that it will be an improvement for sentiment analysis based product ranking techniques in recommendation systems since it affect and improves the correctness of rankings. Also, since the emphasized text handling is an improvement for the domain of NLP, it can be considered as a domain contribution as well as a technical contribution.
Development	A considerable amount of work has been put to this research throughout the phases. From Dataset analyzing and synthesizing to making a predictor model which will then be used with several other core components has been developed with experimentations as well as with a clear goal in mind. A user interface has been developed to demonstrate the core functionalities of the implementation.
Improvements	Through evaluation as well as experimentations during the development process, potential limitations have been identified which can be addressed in future work to improve upon this research.

Table 28: Self evaluation

9.4. Evaluation Results

The implementation was evaluated by Mr. Tharindu Gimhana (Software Engineer at Arimac) and Mr. Nayana Weerasekara (Software Engineer at Arimac) who are experts in the technical field of this research. The opinion of these experts gave ideas on future improvements as well as on the current implementation. Also, evaluated that this research was a good contribution on the technical side since it addresses a very overlooked problem which can be also considered as a domain contribution as well.

9.5. Limitations Of Evaluation

Since there were no existing implemented systems which tried to address the problem of this research, the implemented system was evaluated with simulations of other possible approaches. The primary data science component which was about pre-processing emphasized informal words has been evaluated with self-implemented approaches of possible other methods which can be used but, one of the methods was found through experimentation. Likewise, emoji consideration when ranking products through sentiment analysis of their reviews has not been mentioned in current studies.

9.6. Evaluation On Functional Requirements

ID	Requirement	Priority	Evaluation
FR1	Users must be able to provide the required product recommendations with the reviews so that they can view the ranked products based on sentiment scores of product reviews.	M	Implemented
FR2	Must be able to rank a set of product recommendations based on the calculated sentiment scores of reviews.	M	Implemented
FR3	The system should be able to calculate sentiment scores for product reviews which have emojis and text which may also contain emphasized words to get an overall sentiment score.	M	Implemented
FR4	Preprocess emphasized words in a text content to get the text with actual words instead of emphasized	M	Implemented

	words.		
FR5	Admins should be able to modify system parameters to perform as needed.	C	Not implemented

Table 29: Functional requirements evaluation

9.7. Evaluation On Non-Functional Requirements

ID	Requirement	Priority	Evaluation
NFR1	Performance	S	Completed - Can be improved with reducing memory usage.
NFR2	Scalability	C	Partially Completed - Can handle multiple ranking tasks but, will not be efficient enough to be function as an independent ranking system.
NFR3	Extensibility	S	Completed
NFR4	Quality	C	Completed

Table 30: Non-Functional requirements evaluation

9.8. Chapter Summary

This chapter described the process of evaluation which was conducted on the implementation. It described the evaluation criteria, self-evaluation as well as the overall results obtained from the evaluation process.

10. CONCLUSION

10.1. Chapter Overview

This chapter is about the overall research conclusion with the analysis of research aims and objectives achievements as well as learning outcomes. It also describes limitations and future work to improve upon the research.

10.2. Achievements Of Research Aims & Objectives

Achievement Of Research Aim

The aim of this research is to design, develop, and evaluate a solution which considers the sentiment of emojis and text when ranking products while also being able to correctly preprocess emphasized informal text.

The aim was accomplished by designing, developing, and evaluating a solution that can rank a set of products based on the sentiment of their reviews by considering emojis and text while also correctly preprocessing emphasized informal words within the text to improve the overall process of ranking and sentiment analysis.

Achievement Of Research Objectives

Research Objectives	Description	Status
Literature Review	RO1: Identify existing methods and their limitations in product ranking and recommendation. RO2: Study how current recommendation systems incorporate sentiment analysis. RO3: Identify ways to interpret emojis and emphasize words in a text content. RO4: Conduct a study on how emoji and emphasized words interpretation can affect sentiment analysis.	Completed
Requirement Elicitation	RO1: Gather requirements of product ranking techniques which are used in recommendation systems. RO2: Collect insights and ideas from technology and domain experts.	Completed
Design	RO1: Improve the algorithm for estimating sentiment polarity based on the text content and the emojis in a review. RO2: Design a data pre-process mechanism to obtain a tokenized data set by separating text, emphasized text, and emoji content. RO3: Design a ranking method which can analyze text and emoji data to rank the related items.	Completed

	RO4: Design a machine learning based solution to preprocess emphasized informal texts.	
Implementation	RO1: Develop a ranking method which can analyze the sentiment of text, emphasized text and emoji content to improve the ranking of items. RO2: Develop a machine learning based solution to correctly preprocess emphasized informal texts.	Completed
Testing and Evaluation	RO1: Make an adequate test plan and perform relevant unit, integration, and functional tests. RO2: Evaluate how the developed ranking method improved over existing ones which doesn't make use of emoji content to rank the items. RO3: Evaluate how the machine learning based preprocessing technique cleans emphasized informal text in comparison to existing methods.	Partially completed

Table 31: Achievements of research objectives

10.3. Utilization Of Knowledge From The Course

Modules	Utilized Knowledge
Programming Principles I & II Object Oriented Programming	Foundational knowledge to design and develop a system with proper programming skills and coding concepts.
Web Design and Development Client Server Architecture Server-Side & Advanced Server-Side Web Development	Core web development skills and advanced server-side knowledge gained from these modules was helpful while building a demonstration web app to showcase the system.
Software Development Group Project	Gained the basic and initial knowledge on research & planning a system to be developed.

Industrial Placement	Applied best practices, new concepts and utilized tools which were got to know of during the placement.

Table 32: Course knowledge utilization

10.4. Use Of Existing Skills

- **Server-Side Development** - Worked with latest backend technologies and applied best practices throughout the placement which was completed at Arimac Lanka.

10.5. Use Of New Skills

- **Machine Learning** - Learned basics machine learning through online resources and official documentations of tools such as sci-kit learn.
- **UI/UX Design** - UI/UX best practices were improved by learning through online resources as well as while researching currently existing product ranking user interfaces.

10.6. Achievements Of Learning Outcomes

Learning	**Learning Outcome ID**
Applicable methodologies and tools were selected, justified, and applied for the problem which was going to be solved.	LO1
Relevant requirements were gathered and analyzed using literature review and surveys.	LO3
Current research related to the problem were thoroughly reviewed and evaluated in order to find existing gaps and possible approaches as well as evaluation techniques.	LO4, LO6, LO1
Learned relevant technical skills along with new techniques on performing certain operations in python programming language. Discussed with the supervisor occasionally and improved the solution and the documentation as suggested.	LO5
Implemented a system which solves the addressing problem along with a demonstration interface to view and evaluate the developed solution.	LO7

Every phase of the research has been thoroughly documented and evaluated the implemented system as much as possible in order to provide a good overall document.	LO8

Table 33: Achievements of learning outcomes

10.7. Problems And Challenges Faced

Problems & Challenges	Mitigation
No clear approach on providing a solution for the research gap.	Researched existing works thoroughly and gained knowledge through online resources to develop a proper solution.
The required emphasized words dataset was not available.	Generated manually after analyzing which words are most likely to be emphasized through several review datasets.
Not able to find existing implementations to benchmarking the system to showcase the improvements.	Simulated the existing system in order to compare the proposed solution and benchmark it.

Table 34: Problems & challenges faced

10.8. Deviations

The prototype featured a solution to get the actual sentiment score on emphasized informal words. However, this approach negates the contextual understanding of sentiment analysis models because the sentiment score of these words needs to be combined separately with the sentiment score of the text which contained it. Doing so will result in incorrect sentiment score since they are considered as two separate text content instead of a single text content.

Hence, the machine learning based text classification approach has been chosen to replace the emphasized word with the actual word from the text because that way, the contextual understanding won’t be neglected and will provide correct results in sentiment analysis as well as product ranking.

The asynchronous or concurrency behavior for the backend application was not implemented due to time constraints. The important functional requirements were prioritized.

10.9. Limitations Of The Research

- The emphasized text classification model is currently based on a very limited dataset since it's synthesized. Because of this, the model may not be able to predict the actual words from some emphasized words.
- The synthesized dataset belongs to the ecommerce domain. Other domain specific emphasized words will not be predicted correctly.
- Removes the products with a lower review count and balances the reviews for the selected product to make the system fair.

10.10. Future Enhancements

- Synthesize a dataset with a proper amount of emphasized words which belong to several domains so that a better text prediction model can be trained.
- Identify and exclude words such as "XXL" from the text classification task since these words don't hold any opinion value and may be more inclined towards product features.
- When ranking, the reviews are balanced for each product. Instead of this approach, it will be better to consider all the reviews per product to solve the issue of biased sentiment analysis.
- Context awareness of emojis can be implemented since reviews with sarcastic text are very common in the NLP domain. It will improve the sentiment analysis process as well.
- The product ranking method can be improved with asynchronous or concurrent programming.

10.11. Achievement Of The Contribution To The Body Of Knowledge

10.11.1. Technical Contribution – NLP

- **Emphasized informal words preprocessed using machine learning.**

  A new way to correctly preprocess these noisy texts data which contains repeated letters or intentional misspellings so that it can improve the NLP tasks such as sentiment analysis.

- **Dataset of emphasized words with repeated letters.**

  Since there were no publicly available datasets with emphasized words, the author analyzed several product review datasets and synthesized a dataset of emphasized informal words with repeated letters along with their actual words.

10.11.2. Domain Contribution - Recommendation Systems & NLP

- **Emoji and emphasized informal text consideration in sentiment analysis based product ranking**

  Improve the products ranking mechanism on recommendation systems with the use of sentiment analysis based on emojis and text which may also contain emphasized informal words.

- **Emphasized informal words handling.**

  Also, the preprocessing of emphasized words with repeated letters or intentional misspellings can be very useful in the field of NLP since these words are commonly used by humans. NLP is one of the key domains when it comes to artificial intelligence. Hence, this research is a valuable contribution to the NLP domain as well.

10.12. Concluding Remarks

This concludes the research documentation which is primarily about emoji and text sentiment-based product ranking method which can be used in recommendation systems to improve their ranking results. This research also introduces a new way of preprocessing emphasized informal words with repeating letters or intentional misspellings, so that these words are considered when performing tasks such as sentiment analysis since these types of words are common in the field of NLP and hold valuable information within them.

REFERENCES

Agarwal, A. et al. (2011). Sentiment Analysis of Twitter Data. *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, 30–38.

Bi, J.-W., Liu, Y. and Fan, Z.-P. (2019a). Representing sentiment analysis results of online reviews using interval type-2 fuzzy numbers and its application to product ranking. *Information Sciences*, 504, 293–307. Available from https://doi.org/10.1016/j.ins.2019.07.025.

Bi, J.-W., Liu, Y. and Fan, Z.-P. (2019b). Representing sentiment analysis results of online reviews using interval type-2 fuzzy numbers and its application to product ranking. *Information Sciences,* 504, 293–307. Available from https://doi.org/10.1016/j.ins.2019.07.025.

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

Dang, C.N., Moreno-García, M.N. and Prieta, F.D. la. (2021). An Approach to Integrating Sentiment Analysis into Recommender Systems. *Sensors*, 21 (16), 5666. Available from https://doi.org/10.3390/s21165666.

Das, D., Sahoo, L. and Datta, S. (2017). A Survey on Recommendation System. *International Journal of Computer Applications*, 160 (7), 6–10. Available from https://doi.org/10.5120/ijca2017913081.

Gasmi, S., Bouhadada, T. and Benmachiche, A. (2020). Survey on Recommendation Systems. *Proceedings of the 10th International Conference on Information Systems and Technologies*. 4 June 2020. Lecce Italy: ACM, 1–7. Available from https://doi.org/10.1145/3447568.3448518 [Accessed 2 November 2022].

Giachanou, A. and Crestani, F. (2017). Like It or Not: A Survey of Twitter Sentiment Analysis Methods. *ACM Computing Surveys*, 49 (2), 1–41. Available from https://doi.org/10.1145/2938640.

Guo, C., Du, Z. and Kou, X. (2018). Products Ranking Through Aspect-Based Sentiment Analysis of Online Heterogeneous Reviews. *Journal of Systems Science and Systems Engineering*, 27 (5), 542–558. Available from https://doi.org/10.1007/s11518-018-5388-2.

Heidary Dahooie, J. et al. (2021). An intuitionistic fuzzy data-driven product ranking model using sentiment analysis and multi-criteria decision-making. *Technological Forecasting and Social Change*, 173, 121158. Available from https://doi.org/10.1016/j.techfore.2021.121158.

K, R. and S, M.V. (2022). An Approach for Recommender System Based on Multilevel Sentiment Analysis Using Hybrid Deep Learning Models. *2022 8th International Conference on Smart Structures and Systems (ICSSS)*. 21 April 2022. Chennai, India: IEEE, 01–06. Available from https://doi.org/10.1109/ICSSS54381.2022.9782172 [Accessed 26 January 2023].

Kowsari et al. (2019). Text Classification Algorithms: A Survey. *Information*, 10 (4), 150. Available from https://doi.org/10.3390/info10040150.

Li, H. et al. (2018). Spelling Error Correction Using a Nested RNN Model and Pseudo Training Data. Available from http://arxiv.org/abs/1811.00238 [Accessed 7 May 2023].

Liu, Y., Bi, J.-W. and Fan, Z.-P. (2017). A Method for Ranking Products Through Online Reviews Based on Sentiment Classification and Interval-Valued Intuitionistic Fuzzy TOPSIS. *International Journal of Information Technology & Decision Making*, 16 (06), 1497–1522. Available from https://doi.org/10.1142/S021962201750033X.

Nuspell. (no date). Available from https://nuspell.github.io/.

Sen, P.C., Hajra, M. and Ghosh, M. (2020). Supervised Classification Algorithms in Machine Learning: A Survey and Review. In: Mandal, J.K. and Bhattacharya, D. (eds.). *Emerging Technology in Modelling and Graphics*. Advances in Intelligent Systems and Computing. Singapore: Springer Singapore, 99–111. Available from https://doi.org/10.1007/978-981-13-7403-6_11 [Accessed 8 May 2023].

Tejwani, R. (2014). Sentiment Analysis: A Survey. Available from https://doi.org/10.48550/ARXIV.1405.2584 [Accessed 1 April 2023].

Tian, P. et al. (2009). Research of Product Ranking Technology Based on Opinion Mining. *2009 Second International Conference on Intelligent Computation Technology and Automation*. 2009. Changsha, Hunan, China: IEEE, 239–243. Available from https://doi.org/10.1109/ICICTA.2009.773 [Accessed 27 January 2023].

Wang, W., Wang, H. and Song, Y. (2017). Ranking product aspects through sentiment analysis of online reviews. *Journal of Experimental & Theoretical Artificial Intelligence*, 29 (2), 227–246. Available from https://doi.org/10.1080/0952813X.2015.1132270.

Wankhade, M., Rao, A.C.S. and Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55 (7), 5731–5780. Available from https://doi.org/10.1007/s10462-022-10144-1.

Wu, C. and Zhang, D. (2019). Ranking products with IF-based sentiment word framework and TODIM method. *Kybernetes*, 48 (5), 990–1010. Available from https://doi.org/10.1108/K-01-2018-0029.

Yoo, B. and Rayz, J.T. (2021). Understanding Emojis for Sentiment Analysis. *The International FLAIRS Conference Proceedings*, 34 (1). Available from https://doi.org/10.32473/flairs.v34i1.128562 [Accessed 7 November 2022].

Zhang, D., Wu, C. and Liu, J. (2020). Ranking products with online reviews: A novel method based on hesitant fuzzy set and sentiment word framework. *Journal of the Operational Research Society*, 71 (3), 528–542. Available from https://doi.org/10.1080/01605682.2018.1557021.

APPENDIX

APPENDIX 1 - Implementation

Emphasized Words Identification By Review Datasets Analysis

Figure 1: Amazon Reviews Dataset Analysis

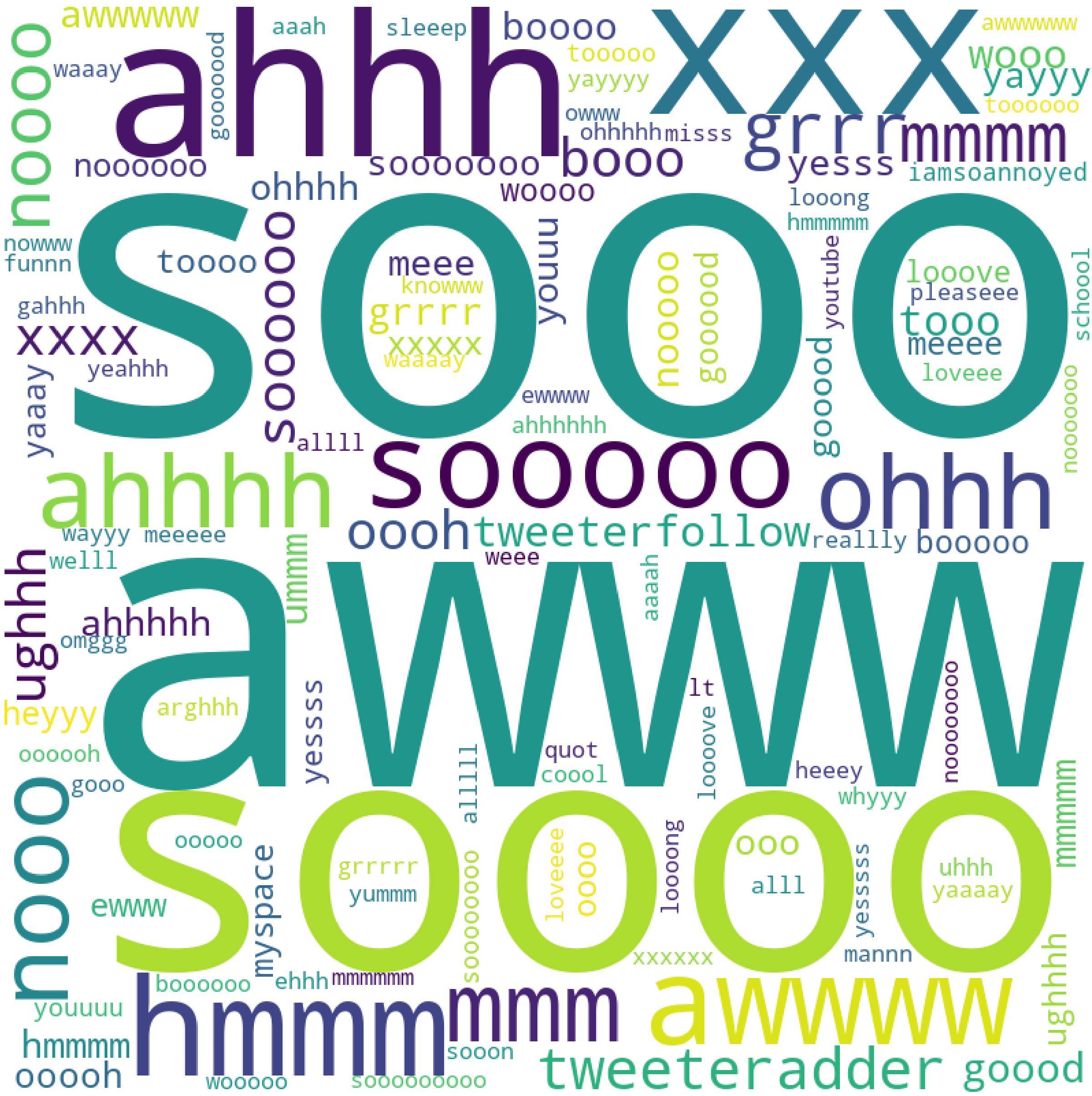


Figure 2: Twitter Sentiment140 Dataset Analysis

Figure 3: Amazon US – Beauty Products Dataset Analysis

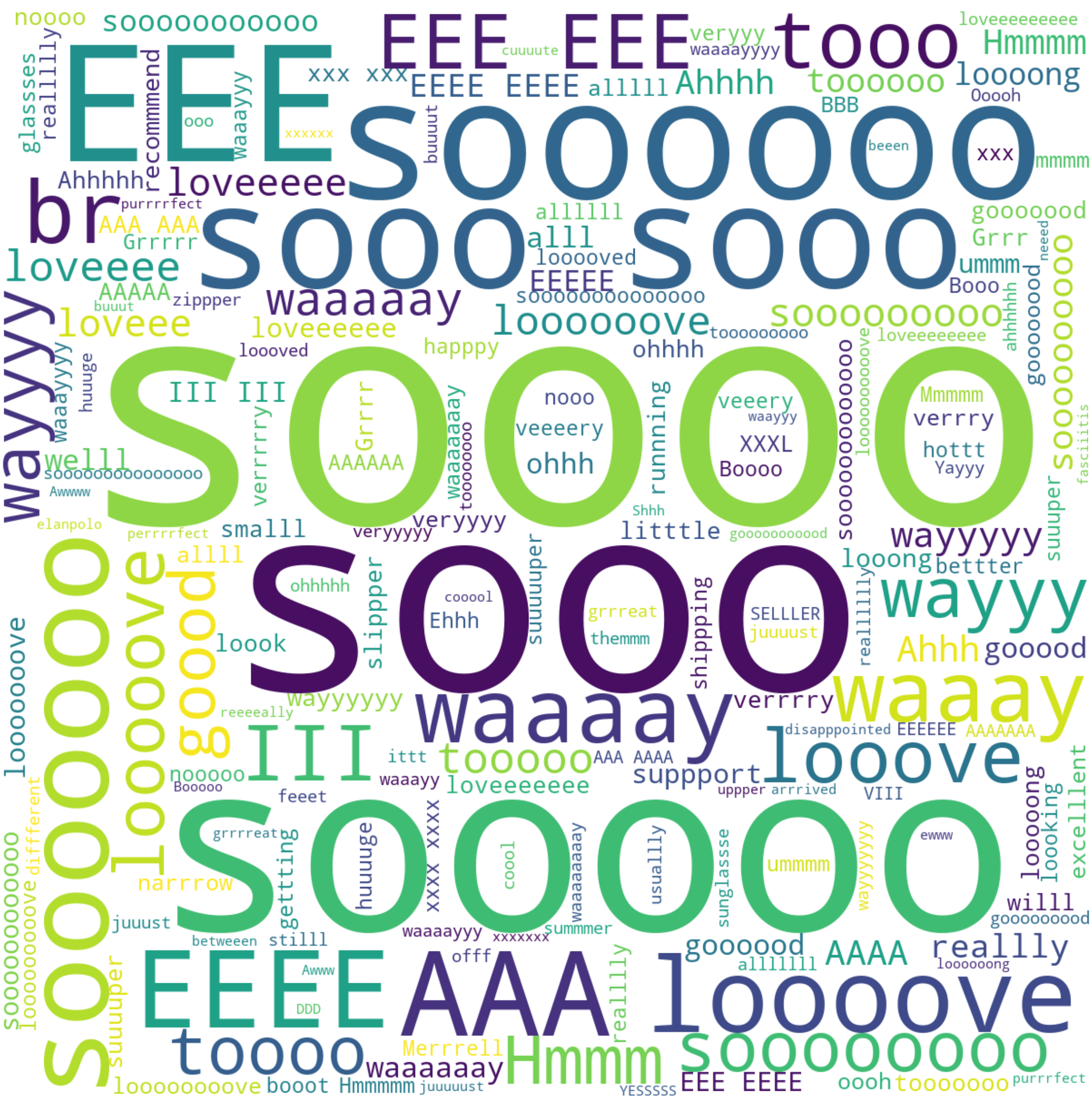


Figure 3: Amazon US – Shoe Products Dataset Analysis

Figure 3: Amazon US – Electronic Products Dataset Analysis