DE-42 (DC & SE)

NUST COLLEGE OF
ELECTRICAL AND MECHANICAL ENGINEERING

CUSTOMER PERSPECTIVE IN E-COMMERCE: SENTIMENT ANALYSIS OF DARAZ REVIEWS

A PROJECT REPORT

DE-42 (DC & SE)

Submitted by

ASC MUHAMMAD TAHIR AMEER

PC MUHAMMAD BILAL KHAN

NS MUHAMMAD TAIMOOR

BACHELORS

IN

COMPUTER ENGINEERING

YEAR

2024

PROJECT SUPERVISOR

DR. ARSLAN SHAUKAT

DEPARTMENT OF COMPUTER & SOFTWARE ENGINEERING
COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,
ISLAMABAD, PAKISTAN

YEAR 2024

Certification

This is to certify that Muhammad Tahir Ameer (359081), Muhammad Bilal Khan (359071) and Muhammad Taimoor (355134) have successfully completed the final project Customer Perceptive in E-Commerce: Sentiment Analysis of DARAZ Reviews, at the National University of Science and Technology, to fulfill the partial requirement of the degree Computer Engineering.

Signature of Project Supervisor
Dr. Arslan Shaukat
Asscociate Professor

Copyright Statement

Sustainable Development Goals (SDGs)

SDG No	Description of SDG	SDG No	Description of SDG
SDG 1	No Poverty	SDG 9 ✓	Industry, Innovation, and Infrastructure
SDG 2	Zero Hunger	SDG 10	Reduced Inequalities
SDG 3	Good Health and Well Being	SDG 11	Sustainable Cities and Communities
SDG 4	Quality Education	SDG 12	Responsible Consumption and Production
SDG 5	Gender Equality	SDG 13	Climate Change
SDG 6	Clean Water and Sanitation	SDG 14	Life Below Water
SDG 7	Affordable and Clean Energy	SDG 15	Life on Land
SDG 8	Decent Work and Economic Growth	SDG 16	Peace, Justice and Strong Institutions
		SDG 17	Partnerships for the Goals

1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS
SUSTAINABLE DEVELOPMENT GOALS

Sustainable Development Goals

Complex Engineering Problem

Range of Complex Problem Solving

	Attribute	**Complex Problem**	
1	Range of conflicting requirements	Involve wide-ranging or conflicting technical, engineering and other issues.	
2	Depth of analysis required	Have no obvious solution and require abstract thinking, originality in analysis to formulate suitable models.	**X**
3	Depth of knowledge required	Requires research-based knowledge much of which is at, or informed by, the forefront of the professional discipline and which allows a fundamentals-based, first principles analytical approach.	**X**
4	Familiarity of issues	Involve infrequently encountered issues	**X**
5	Extent of applicable codes	Are outside problems encompassed by standards and codes of practice for professional engineering.	**X**
6	Extent of stakeholder involvement and level of conflicting requirements	Involve diverse groups of stakeholders with widely varying needs.	
7	Consequences	Have significant consequences in a range of contexts.	**X**
8	Interdependence	Are high level problems including many component parts or sub-problems	**X**

Range of Complex Problem Activities

	Attribute	**Complex Activities**	
1	Range of resources	Involve the use of diverse resources (and for this purpose, resources include people, money, equipment, materials, information and technologies).	**X**
2	Level of interaction	Require resolution of significant problems arising from interactions between wide ranging and conflicting technical, engineering or other issues.	**X**
3	Innovation	Involve creative use of engineering principles and research-based knowledge in novel ways.	**X**
4	Consequences to society and the environment	Have significant consequences in a range of contexts, characterized by difficulty of prediction and mitigation.	**X**
5	Familiarity	Can extend beyond previous experiences by applying principles-based approaches.	**X**

Dedicated to

Our final year project is dedicated to our supervisor Dr Arslan Shoukat, whose expertise, patience, and invaluable feedback were essential to its success. This work is also dedicated to the support and encouragement of our **parents**, who have been my greatest source of strength and motivation throughout this process. I would like to express my gratitude to our teachers, whose guidance and knowledge have been instrumental in my academic and personal growth, as well as to our friends and peers for their continuous companionship and support throughout this academic and personal process. All the wonderful people in our life contributed to this accomplishment through their collective efforts and encouragement.

Acknowledgment

First of all, Alhamdulillah, that our FYP is finally made and all Thanks to Allah for giving us the strength and moral to keep pushing forward and helping us on each and every step of the way. Secondly, we would like to offer heartily thanks to our supervisor, Dr. Arslan Shaukat, who helped us a lot, tremendously, on each and every single issue, whose help and guidance became a source of strong determination for us. Thank you, sir, you played a great role in our lives, one that we can never forget. And lastly, we would like to thank our parents and friends, without whose unimaginable support and constant motivation, we might not have been able to complete our final year project. They played an unparalleled role throughout our journey and we are eternally thankful to them. Their constant support motivated us to do more than we ever realized, and they inspired new hope in us, when we found none in ourselves.

Abstract

In recent times, there has been a huge boom in the trend of online shopping. This has led to an increase of data on such sites, especially in the form of reviews left by customers. Depending on the country these websites are based in, these reviews can be in many languages. Roman Urdu, a language spoken mainly in the subcontinent, is an example of one such language. While lots of work has been done in the field of sentiment analysis on popular languages like English, German and French, the same claim cannot be made for Roman Urdu. This project aims to address this gap. The goal is to explore natural language processing techniques to accurately classify Roman Urdu text. For this purpose, reviews from the popular Pakistani e-commerce website Daraz.pk are used. The data was first preprocessed to remove stop words and emojis. Then, to classify this data, a DistilBERT model was finetuned on more than 10,000 reviews from Daraz.pk, achieving an accuracy of 85 percent. Moreover, a web interface will be provided where the user can paste a link of the product they want to buy, and generate visualizations based upon which they can decide on buying the product or not. The potential users of this product are customers of Daraz.pk who want a more robust system to decide whether a product should be bought or not.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Sentiment analysis is the process of using natural language processing techniques to classify text as positive, negative, or neutral. In recent years, there has been a huge boom in online shopping. This has led to the creation of lots of data on these customers, especially in the form of reviews left on the products on the websites by customers visiting these sites. Depending on the country the website is based in, the reviews come in many languages. One such language is Roman Urdu, used predominantly in the subcontinent. There is a lot of work in the field of sentiment analysis that has been done in popular languages like English, German and French. Roman Urdu, in comparison, does not have much work done in this field. This project aims to explore sentiment analysis techniques to accurately classify Roman Urdu text. For this purpose, we are using reviews from Daraz.pk. Daraz.pk is a Pakistani e-commerce website with a rich collection of reviews on a vast array of products. A lot of these reviews are in Roman Urdu. An online platform will be made in the form of a web application to achieve this task which will enable the users to sign in and paste a link of the product they want to buy. A web crawler will extract all the relevant reviews from the product link. All these reviews will then be sent to a machine learning model which will process these reviews and generate insights in the form of a pie chart or a bar chart which will guide the user to decide on if they want to buy the product or not. The project's primary goal is to enable users to be able to make better decisions regarding what product to buy. Natural language processing language techniques are used to that end as they are more robust compared to manually looking at the ratings of the products the user wants to buy from Daraz.pk.

1.1 Motivation

In the present day, customer reviews are a goldmine of insights for businesses and researchers. This feedback is capable of guiding improvements in products already on the market, enhance customer service and give the product owners clues about how to market their product. These reviews are written in many languages depending on the country and region they are located in. In Pakistan, Roman Urdu is the main language used on online platforms. E-commerce stores like Daraz.pk are no exception to this rule, where a lot of the reviews on products use Roman Urdu. Roman Urdu is a resource-poor language[1]. Despite its wide use, it is quite underrepresented in the field of natural language processing compared to more popular languages like English, German and French. As such, we believe that exploring techniques to accurately perform sentiment analysis on Roman Urdu will be very beneficial. The limited research that has been performed in the field of sentiment analysis of Roman Urdu reviews is in academic spheres, with little to no practical applications. Various algorithms and machine learning models have been tested to see if they produce satisfactory results on Roman Urdu text. However, a large gap exists between theoretical research and real-world applications. It is a gap that we aim to address in this project.

1.2 Scope

The project‘s aim is to develop an online platform that will allow potential buyers of a product on Daraz.pk to sign in and paste the link of the product they are interested in buying. The website will crawl the given product link to extract relevant reviews. Once the reviews are extracted, they will be sent to a machine learning model which will classify the reviews as positive, negative, or neutral. The ratio of these ratings will be provided to the user in the form of a pie chart. Users will also be able to view their search history on a separate page. This will give a list of the products the users searched while using our online platform, as well as their positive, negative and neutral score, which will be displayed as percentages.

The scope of the project can be defined in terms of the following objectives: • Finetuning a transformer-based machine learning model to perform sentiment analysis on Roman Urdu. • Development of an online platform that will allow the user to sign in, get all the reviews on a product link, and generate insights from those reviews. • Creation of a dataset based upon reviews from Daraz.pk. • Developing a database that will store user search history, and will allow a user to view all the products they searched while using the platform.

1.3 Aims

The goal of this project is to create a sentiment analysis tool that is accurate and reliable and is especially made for the Daraz online store. Using sophisticated natural language processing (NLP) models that have been refined on a dataset of customer evaluations from Daraz, this tool will categorize reviews into three categories: neutral, negative, and positive sentiment. In order to improve product offerings, boost customer satisfaction, and guide strategic business decisions throughout the Daraz ecosystem, the tool will enable real-time analysis and deliver actionable information.

1.4 Objectives

1.4.1 Data Collection and Labeling

It makes sure that pre-trained language models like DistilBERT and Multilingual BERT understands the unique linguistic complexities and contextual attitudes. They communicate in Daraz reviews by fine-tuning on a labeled dataset that is collected earlier from Daraz.

1.4.2 Model Fine-Tuning

It makes sure that pre-trained language models like DistilBERT and Multilingual BERT understands the unique linguistic complexities and contextual attitudes. They communicate in Daraz reviews by fine-tuning on a labeled dataset that is collected earlier from Daraz.

1.4.3 Real-Time Data Acquisition

Furthermore, an intelligent web scraper is made to automatically collect customer reviews from Daraz product links in real-time. It is then connected with a sentiment analysis website to provide reliable and continuous data collection.

1.4.4 Sentiment Classification API

Additionally, the DistilBert model that is working reliably divide customer reviews into three categories: positive, negative, and neutral. This model is converted to API to process and categorize reviews more quickly.

1.4.5 Data Storage and Management

A dependable database system is designed and implemented to store sentiment analysis data as well as long-term research findings on customer sentiment patterns.

1.4.6 Insights and Reporting

subsequently, a visualization website is created to support accurate choices and strategic planning. Merchants, customers, and the Daraz platform may all access sentiment analysis data and comprehensive reports and visualizations with the use of user-friendly interfaces.

The objective of the project is to create a powerful sentiment analysis tool that enhances the overall Daraz user experience and contributes to the platform's growth and success.

1.5 Outcomes

The effective deployment of the sentiment analysis tool on the Daraz e-commerce platform is expected to produce multiple noteworthy results, offering major benefits to multiple stakeholders. First and foremost, customers will have better shopping experiences since they will have access to condensed sentiment information about products, which will enable them to base their judgments on a larger body of feedback. Increased platform trust because of this transparency will boost client happiness and loyalty. Conversely, sellers will possess insightful knowledge about what customers think of their goods, allowing them to spot opportunities for development and take proactive measures to resolve complaints. By facilitating product modifications that better satisfy customer expectations and preferences, this feedback analysis will raise the overall quality of the product. By recognizing consumer preferences and market trends over time, the sentiment analysis technology will help Daraz with its marketing and strategic planning by offering strategic insights. Furthermore, examining general sentiment trends will point out opportunities for platform improvement, such better customer service and UI improvements. Technically speaking, very accurate sentiment categorization may be achieved by fine-tuning sophisticated NLP models like DistilBERT and Multilingual BERT on a dataset unique to Daraz. This will establish a standard for future NLP applications in e-commerce. By incorporating web scraping to collect data in real-time, the sentiment analysis tool will remain relevant and valuable by offering the most recent findings.

Moreover, longitudinal studies will be made possible by the creation of a well-organized database to hold sentiment analysis data, which will aid in monitoring changes in customer attitude and behavior over time. This repository will be an invaluable tool for upcoming studies and advancements, promoting the ongoing refinement of sentiment analysis methods and applications. All things considered, the sentiment analysis tool will yield useful information that can be used to improve product offerings, raise consumer contentment, and guide strategic choices—all of which will contribute to the expansion and prosperity of the Daraz ecosystem.

1.6 Report Organization

The organization of the thesis is as follows:

1.6.1 Introduction

A summary of the project, including its history, goals, and relevance, is given in this section. It draws attention to the problem statement, highlighting the benefits expected for different stakeholders and the necessity of a sentiment analysis tool specifically designed for the Daraz e-commerce platform.

1.6.2 Literature Review

The literature study looks at methods and tools for sentiment analysis that are currently in use, with an emphasis on natural language processing (NLP) approaches. Giving a theoretical basis for the project's methodology, it analyzes related work on fine-tuning pre-trained models such as DistilBERT and Multilingual BERT for datasets.

1.6.3 Methodology

This section describes the project's approach in detail, including the procedures for labeling and gathering data. It describes the processes used to optimize pre-trained models, such as the choice of DistilBERT and Multilingual BERT and the optimization methods applied. It also describes how real-time data is obtained via web scraping and how the sentiment analysis tool is integrated with it.

1.6.4 Implementation

Technical information regarding the sentiment analysis tool's development is provided in the implementation section. It explains the tool's design, how to integrate an API to process reviews, and how to store and arrange studied data in databases.

1.6.5 Results and Analysis

The sentiment analysis tool's accuracy rates and performance data are shown in this section. Key insights and patterns found in customer evaluations on the Daraz platform are highlighted, along with a study of sentiment trends obtained from the data collected.

Chapter 2

Literature Review

The purpose of literature review is to provide a thorough summary of existing research and theoretical work related to sentimental analysis. In our case, Literature review plays an important role because we have to search for the model which we are going to use in our FYP. So we start searching for the papers which will help us in our FYP. The first paper we find related to our FYP is from our own NUST University.

The paper explores sentiment analysis in Roman Urdu, a resource-poor language, leveraging advanced deep learning models like DistilBERT to address the challenge of limited data availability [1]. Roman Urdu poses challenges like variations in spellings, lack of standard writing system, morphological richness, and absence of capitalization, making sentiment analysis complex. Logistic Regression, Naive Bayes, DistilBERT, were compared for sentiment analysis on a dataset of 21k labeled Roman Urdu sentences. DistilBERT achieved 85 percent accuracy with only 2 epochs, outperforming other models. Sentiment analysis is crucial for organizations to recognize consumer sentiments and upgrade product/service condition based on feedback shared online.

After this studying this paper, the things are getting clear to us. But we continued our literature review because we have to search a model that gives us the accuracy nearly 90 percent.

The second paper emphasizes the significant role of sentiment analysis in discerning audience sentiments regarding a specific topic or product.[2]. They follow the standard mechanisms of Machine learning and natural language processing. They obtained the dataset from Kaggle, contains the reviews of distinct products in Roman Urdu text that have been posted on the Daraz E-Commerce website. Daraz is a prominent online shopping platform in South Asia, initially starting as an online fashion store in Pakistan before being acquired by Alibaba Group to expand its presence in the region. The platform allows users to leave reviews on products purchased from vendors, with reviews being predominantly in English globally but in Roman Urdu in Pakistan to cater to the local population's language preference. Roman Urdu is essentially Urdu language written using English alphabets, making it easier for Urdu speakers to read and understand the reviews on Daraz. The use of Roman Urdu on Daraz in Pakistan reflects the platform's adaptation to local linguistic preferences, enhancing user experience and engagement After the data is collected and then preprocessed. Then, by using different approaches of machine learning classifiers in python dataset is analyzed. Logistics Regression outperforms with an accuracy of 75%.

Third paper revolves around the sentiment analysis involves studying attitudes, opinions, and sentiments towards various subjects using computational methods[3]. It is crucial for companies and customers to make informed decisions by analyzing social media text for opinions and feedback. The rise of social media has led to a vast amount of text in languages like code-mixed Roman Urdu and English, which poses challenges for sentiment analysis due to its informal nature Existing sentiment analysis techniques struggle with the nuances of code-mixed languages, necessitating the use of advanced deep learning models for improved accuracy The paper aims to conduct sentiment analysis on code-mixed Roman Urdu and English social media text using up-to-date deep learning models like Multilingual BERT (mBERT) and XLM-RoBERTa (XLM) without relying on lexical normalization or language dictionaries The XLM-R model, with enhanced hyperparameters, outperformed the mBERT model in sentiment analysis of code-mixed Roman Urdu and English social media text, achieving an F1 score of 71%.

The fourth research focuses on sentiment analysis of reviews in Roman Urdu, creating a model to classify polarity in Roman Urdu text[4]. A dataset of 24,000 manually annotated reviews was created by scraping reviews from YouTube for twenty selected songs. The need for a target dataset on Roman Urdu reviews led to the development of the dataset DRU for sentiment analysis, collected from YouTube reviews. The methodology involved dataset collection from YouTube, manual annotation, and sentiment analysis on Roman Urdu reviews . Existing works on Roman Urdu sentiment analysis were discussed, highlighting a study on 1,600 hotel reviews . The paper is structured into sections discussing literature review, methodology, corpus generation, experimental results, and conclusions for possible future enhancements.

Last paper in literature review demonstrates the efficiency of machine learning models in sentiment analysis of Roman Urdu text, highlighting the importance of considering different languages in sentiment mining research[5]. Sentiment analysis plays a crucial role in various online platforms like social media networks, marketing websites, and communication forums. It involves analyzing comments and reviews to extract subjective information about user attitudes towards different topics. The study focuses on sentiment analysis of Roman Urdu text using machine learning models. A dataset of 3000 hotel reviews in Roman Urdu was collected and preprocessed for analysis. Unique machine learning classifiers were applied to predict the polarity of Roman Urdu text models. Logistic regression and Support Vector Machine (SVM) showed superior performance in terms of accuracy, recall, precision, and F-measure The research aims to contribute to the understanding of sentiment analysis in resource-constrained languages like Roman Urdu, which has received limited attention compared to other languages like English, Urdu, and Arabic.

So, after reading all these research papers, we find two models for our FYP which exactly collide with our requirements. These two models are DistilBERT and multilingual BERT (mBERT). After working on both these models with the same dataset, we concluded, distilBERT was working much better than multilingual BERT (mBERT) for several reasons. DistilBERT was smaller in size, hence it has more speed to process the reviews. One more reason was that it has more accuracy than multilingual BERT.

Chapter 3

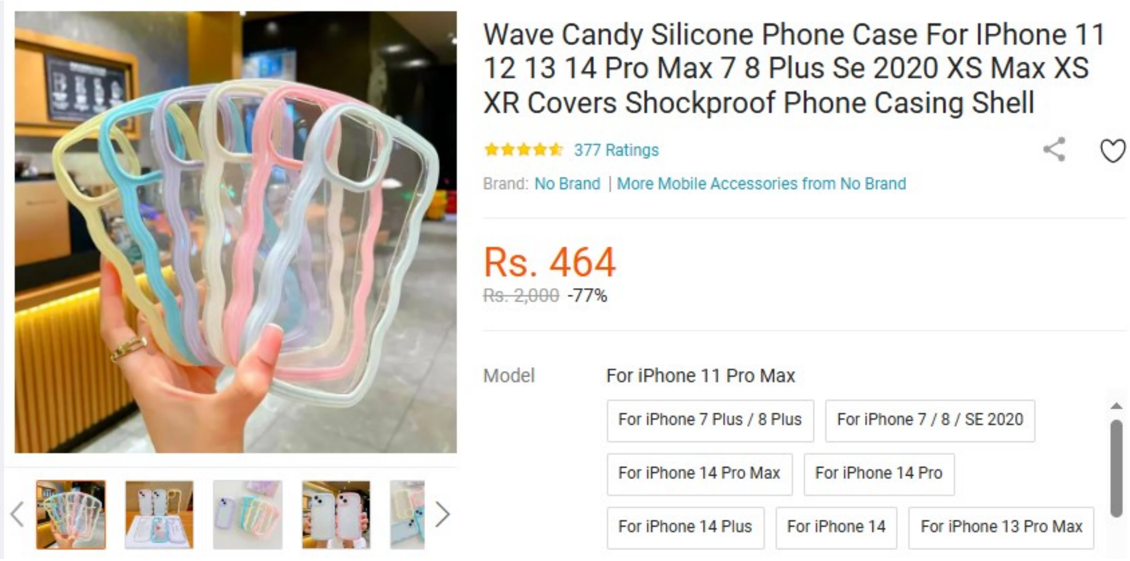
Related Work

3.1 Related Products

The proposed project can be comprehensively classified into the following categories from the development and product perspective:

- Creation of a dataset based upon reviews from the Pakistani e-commerce website Daraz.pk.
- Fine tuning a transformer-based model on the created dataset.
- Development of an online platform that will allow users to paste a product link and get insights on whether the product should be bought or not.
- Development of a database that will store user history. A user will be able to view all the products they searched along with their positive, negative and neutral scores while using our online platform.

While work has been done to accurately perform sentiment analysis on Roman Urdu, this has mostly happened in academia. So far, there have been no practical applications for that research. Currently, no online e-commerce platform utilizes the power of machine learning to perform sentiment analysis. They usually rely on star ratings left by customers on their products. The problem with such ratings is that it might lead to false positives (bad review, good star rating) or false negatives (good review, bad star rating). Using natural language processing techniques to classify such reviews provides a more robust and accurate way of knowing the true sentiment regarding a product and whether it should be bought or not. As it has been already stated, there is no prominent online e-commerce platform in Pakistan that uses natural language processing to give insights to the customer regarding whether a product should be bought or not. They mainly rely on star ratings to inform customers about the general sentiment of a product. The review rating system of Daraz.pk is one such example[6] as shown in Figure 1. Amazon Comprehend is a natural language processing service that uses machine learning to uncover insights and connections in text[7]. It comes with a sentiment analysis API[8]. MonkeyLearn is a no code text analytics platform [5] (MonkeyLearn Inc., 2014). It provides various services related to text classification as well, one of which is sentiment analysis.



Wave Candy Silicone Phone Case For IPhone 11 12 13 14 Pro Max 7 8 Plus Se 2020 XS Max XS XR Covers Shockproof Phone Casing Shell

★★★★★ 377 Ratings

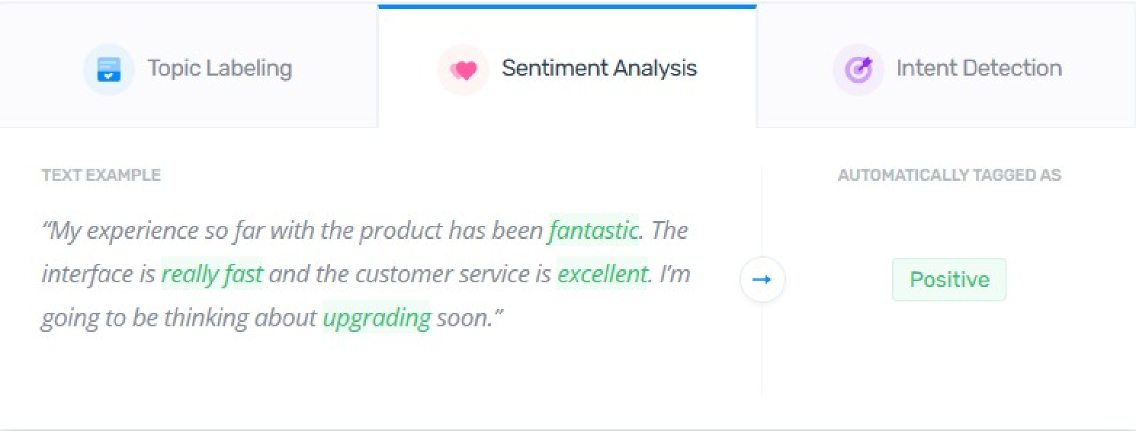
Brand: No Brand | More Mobile Accessories from No Brand

Rs. 464

~~Rs. 2,000~~ -77%

Model For iPhone 11 Pro Max

Figure 3.1: An example of the star rating system utilized by Daraz.pk



"My experience so far with the product has been fantastic. The interface is really fast and the customer service is excellent. I'm going to be thinking about upgrading soon."

Figure 3.2: MonkeyLearn's sentiment analysis service

IBM Watson Natural Language processing is a text analytics service that provides support for over 13 languages[9]. It supports various natural language processing tasks, including sentiment analysis.

IBM Watson Natural Language Understanding
The natural language AI service for advanced text analytics

Figure 3.3: IBM Watson Natural Language Understanding

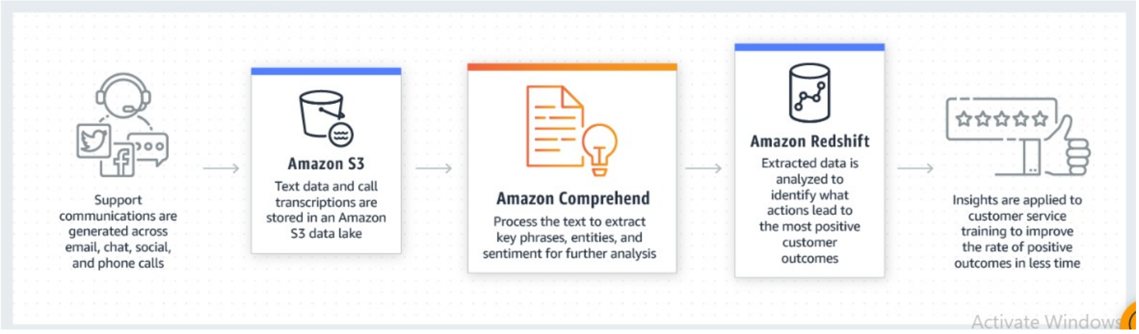

Support communications are generated across email, chat, social, and phone calls
Amazon S3
Text data and call transcriptions are stored in an Amazon S3 data lake
Amazon Comprehend
Process the text to extract key phrases, entities, and sentiment for further analysis
Amazon Redshift
Extracted data is analyzed to identify what actions lead to the most positive customer outcomes
Insights are applied to customer service training to improve the rate of positive outcomes in less time
Activate Windows

Figure 3.4: Working of Amazon Comprehend

Figure 3.4 shows the working of Amazon Comprehend. The first step is extracting data from various social media platforms. These reviews are stored in Amazon S3, which is a storage system utilized by Amazon. This text is then analyzed by Amazon Comprehend, which extracts sentiment from it. These extracted sentiments are then analyzed by Amazon Redshift, which is helpful for telling what actions lead to the most positive customer experience.

Product	Cost	Features	Website
Daraz.pk Star Ratings	-	Tells the overall rating of a product using star ratings	https://www.daraz.pk/
Amazon Comprehend	0.000025$ per 50 million units	Provides a sentiment analysis API. Does not provide support for Roman Urdu text.	https://aws.amazon.com/comprehend/
MonkeyLearn Inc.	299$ /month	Provides text classification services which also includes sentiment analysis.	https://monkeylearn.com/text-classifiers/
Kapiche	2650$/month	Uses artificial intelligence to perform analysis on customer data and automatically generate reports based on that.	https://www.kapiche.com/
Lexalytics	Custom pricing based upon features	Has various text analytics APIs which can be integrated into applications. Includes a sentiment analysis API.	https://www.lexalytics.com/
Repustate	$199 / month	Analyzes customer feedback in various languages and helps gain insight into customer sentiments.	https://www.repustate.com/
Quid	Custom depending on features	Helps market experts and strategists analyze large volumes of data	https://www.quid.com/
IBM Watson Natural Language Understanding	$0.003 per text record	Uses deep learning to extract meaning from unstructured text data. Can be used to extract sentiments.	https://www.ibm.com/products/
Brandwatch	Custom depending on features	Allows businesses to collect reviews, categorize them and use AI to spot insights and generate questions	https://www.brandwatch.com/
Talkwalker Customer Feedback Analytics	$800/month	Allows businesses to monitor customer feedback across all feedback channels and get actionable insights.	https://www.talkwalker.com/products/

Table 3.1: Table for existing technologies

Chapter 4

Development Work

4.1 Data Collection

In Pakistan, e-commerce has grown significantly, and Daraz has become a prominent player in the market. Data is essential to generating insights and ideas in this scenario. Our senior project's goal is to create an extensive dataset from Daraz that can be applied to a variety of tasks, such as product recommendation engines, market analysis, and consumer behavior research. Our goals include compiling extensive data from Daraz's several categories, guaranteeing the accuracy and applicability of the information, and using this dataset in practical contexts to extract valuable insights. Prior research on the gathering of data for e-commerce has emphasized the significance of data in improving the online shopping experience and informing business plans. By examining these research, we hope to advance current understanding and customize our strategy to the particulars of the Pakistani market, as represented by Daraz. This assessment of the literature will serve as the basis for our approach and assist us in addressing any potential difficulties with data collecting and processing.

4.2 Data Labelling

We manually classify the dataset after it was gathered to make sure it was accurate and correct to the uses we had in mind. The data was manually labeled. This improved the dataset's value for a range of analytical activities. Our classification of dataset into three categories—positive, negative, and neutral—was an important part of the labeling process. We have made sure that the dataset is reliable and complete by labeling each review, which makes it a strong tool for creating product recommendation systems, analyzing markets, and researching consumer trends. Previous studies on data collection for e-commerce have shown the value of data in enhancing the online shopping experience and guiding business plans. Precise labeling is essential because it directly affects the performance of analytical tools and machine learning models, which depend on high-quality data to produce insightful results. Our careful approach to data labeling not only helps to achieve the objectives of our project but also advances the field of e-commerce research by offering a trustworthy and well-organized dataset for upcoming research.

4.3 Data Preprocessing

Preprocessing data is an essential stage in machine learning projects, particularly for sentiment analysis and other natural language processing (NLP) applications. A large dataset has been gathered from one of the top e-commerce platforms, Daraz. As could be expected, the data was unclean and needed a lot of preprocessing before it could be used to train sentiment analysis models. The preprocessing procedures use to clean and prepare the data are described in detail below, guaranteeing excellent performance for our refined DistilBERT and Multilingual BERT models.

4.3.1 Removing Stop Words

Stop words are frequently used terms with little meaning that are typically eliminated from datasets in order to minimize their dimensionality. Some instances are "q" ,"k", "ksy," in roman urdu and so on. Despite being necessary for sentence structure, these words don't add anything to the review's tone. We eliminated these stop words by using the Natural Language Toolkit (NLTK) and specially crafted lists designed for the Roman Uedu language. This phase made it easier to concentrate on the words that convey more important sentiment information. Hence, improved performance.

4.3.2 Removing Emojis

Emojis are often used in online evaluations and are a useful tool for expressing emotions. They do, however, also add noise, which can make textual analysis more difficult. In this phase, we identified and eliminated every emoji from the text using regular expressions. Since our models are predominantly text-based, eliminating emoticons helps to preserve consistency in data representation, even though emojis might occasionally provide useful sentiment clues.

4.3.3 Tokenization

The practice of breaking up text into discrete words or tokens is known as tokenization. DistilBERT and Multilingual BERT tokenizers are on reviews, which are unique to the models we selected, for this project. These tokenizers divide the text into words and then format those words so that they may be entered into BERT models.

4.3.3.1 DistilBERT Tokenizer

Using a breakdown of the text, this tokenizer associates each word in the DistilBERT vocabulary with a corresponding token ID. A more compact and lightweight version of BERT, called DistilBERT, keeps most of its features. Special tokens like '[CLS]' for classification and ‘[SEP]’ for sentence separation are also handled by the tokenizer.

4.3.3.2 Multilingual BERT Tokenizer

It is essential to employ a tokenizer[10] that could handle different languages because the reviews were multilingual. While it supports more languages than the DistilBERT tokenizer, the Multilingual BERT[11] tokenizer functions similarly. This is crucial to ensure that sentiment analysis is correct in a variety of linguistic situations, especially for evaluations written in languages other than English.

4.3.4 Lowercasing and Normalization

Every text record has to be change to lowercase following tokenization. By ensuring that terms like "acha" and "Acha" are treated as the same token by the model, this step helps to eliminate redundancy. To manage contractions and frequent misspellings, we also carried out text normalization, which helped to further refine the dataset.

4.3.5 Removing Special Characters

Punctuation, numerals, and other non-alphabetic characters were eliminated along with special characters. These characters can be a distraction and don’t add anything to the sentiment. In order to effectively remove these characters from the text, regular expressions were employed.

4.3.6 Final Preprocessing Steps

Lastly, we made sure that every review was suitably padded or shortened to meet the input size specifications of the model. For batch processing to take place during model training, this homogeneity is required. Following preprocessing, the data was saved in a manner suitable for model input.

4.4 Model Selection

After reading all research papers related to our FYP, we find two models for our FYP which exactly collide with our requirements. These two models are DistilBERT[12] and multilingual BERT (mBERT). After working on both these models with the same dataset, we concluded, dis tilBERT was working much better than multilingual BERT (mBERT) for several reasons. DistilBERT was smaller in size, hence it has more speed to process the reviews. One more reason was that it has more accuracy than multilingual BERT.

4.5 Transformer Based BERT Architecture

After literature review, it was decided to use either DistilBERT, which is a distilled version of the original BERT[13] architecture or BERT multilingual[14]. The NLP models listed above are all variants of the original BERT architecture, and thus use the transformer architecture introduced for the first time in 2017. The transformer architecture consists of encoder layers and decoder layers, both of which are usually six in number[14]. Figure shows a small subset of a transformer consisting of one encoder and one decoder. The working of encoder and decoder is described in subsequent sections:

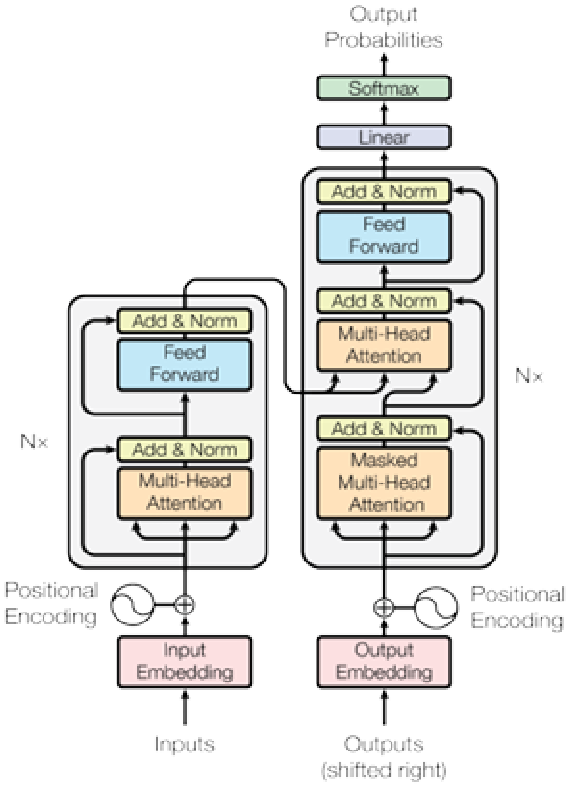


Figure 4.1: The Transformer architecture as described in “Attention is All You Need”

4.5.1 Encoder Layer

The encoder is a stack of 6 identical layers, with each layer containing two sublayers[14]. The first layer is a multi-head self-attention layer and the second layer is a feed forward neural network. Both of these sublayers employ residual connections and layer normalization[14].

4.5.2 Multi-Head Self-Attention Layer

The multi-head self-attention layer in the BERT architecture consists of twelve self-attention heads[15]. These heads work in parallel to operate on queries and keys, as well as values which are sent in as input. The dot product of one query is computed with all keys, and it is scaled by a factor of dk , which is the dimensions of the keys sent in as input[16]. A softmax function is applied to get the weights that will be used on the values.

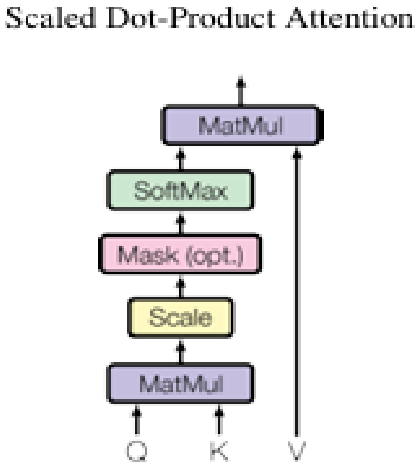


Figure 4.2: Working of the scaled dot product attention employed in the Transformer architecture

A general formula of this attention function[17] is described as below:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4.1}$$

4.5.3 Feed Forward Neural Network

In addition to the attention sublayers, the encoder also uses a feed forward neural network. The feed forward neural network is applied to each position separately. It consists of two linear transformations with a RELU activation in-between:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{4.2}$$

4.5.4 Decoder Layer

The decoder layer, similar to the encoder, is a stack of six identical layers[17]. Just like the encoder, it consists of a multi-head self-attention layer and a feed forward neural network, both of which have been described in the previous section. The only addition is that of a masked multi-head attention layer.

4.5.5 Masked Multi-Head Self-Attention Layer

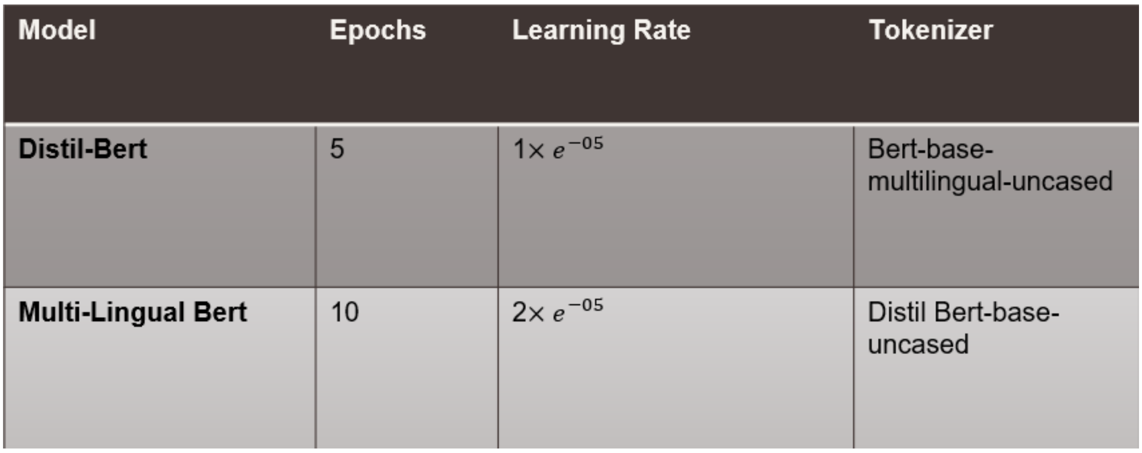
The masked multi-head self-attention layer in a decoder is a modification of the self-attention layer. Masking is done to prevent it from reading subsequent positions, and ensure that only known embedding values are read to predict the future value for a certain position[18].

4.6 Model Training

Following the dataset collection, we used it to train two models of natural language processing: mBERT[19] and DistilBERT[11]. DistilBERT is a quicker and more compact version of BERT that is perfect for real-world applications since it can achieve great performance with less processing power. Given Pakistan's varied linguistic terrain, mBERT, or multilingual BERT, is especially helpful because it can comprehend various languages. We attempted to take advantage of these models' capacities to reliably classify reviews into good, negative, and neutral categories by training them on our manually labeled dataset from Daraz. During the training phase, these models had to be adjusted to our particular dataset's special features and subtleties. This methodology not only increases the models' comprehension and interpretation of user evaluations, but it also raises the general efficacy and accuracy of our consumer behavior analysis and product recommendation systems. Our models now have the capacity to offer more accurate forecasts and in-depth insights thanks to this instruction, which has greatly aided in the accomplishment of our senior project.

4.7 Model Hyperparameters

So below mentioned parameters are used to test these models on our dataset which we collected from the Daraz through web scraping. After a lot of research and literature review on these models we have decided to test these parameters. And these parameters provide us the best results with respect to other parameters that we are using for testing.



Model	Epochs	Learning Rate	Tokenizer
Distil-Bert	5	$1\times e^{-05}$	Bert-base-multilingual-uncased
Multi-Lingual Bert	10	$2\times e^{-05}$	Distil Bert-base-uncased

Figure 4.3: Model Hyperparameters

4.8 Model Accuracy

When evaluating the effectiveness of machine learning algorithms, model accuracy is a crucial parameter, especially for applications like consumer behavior research, market analysis, and product recommendation systems. A high accuracy level means that the model accurately forecasts results or classifies data points according to the real labels or anticipated outcomes. High model accuracy requires a number of critical processes, such as feature selection, algorithm improvement, and data pretreatment. Ensuring model accuracy was critical to our research, particularly with our Daraz hand labeled dataset. Our algorithms' capacity to deliver trustworthy suggestions, perceptive market trends, and precise customer sentiment analysis is strongly impacted by their correctness. We may increase the predictive capacity and optimize the parameters of our models by thoroughly testing and validating them.

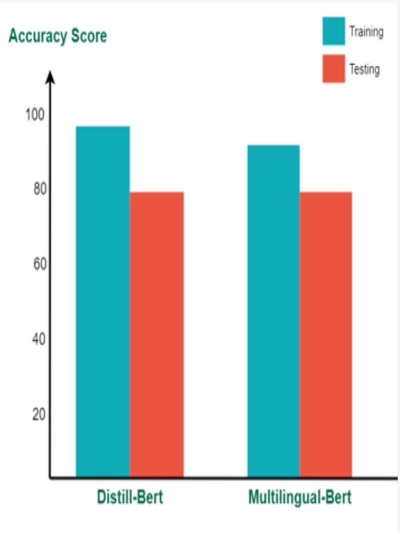


Figure 4.4: Models Accuracy

Sentiment analysis project outcomes shows how well pre-trained language models performed when adjusted using the Daraz review dataset. On the training set, we used DistilBERT to reach an amazing 93% accuracy; however, the accuracy on the test set was somewhat lower at 79%.

However, Multilingual BERT did better on the test set, scoring 80% accuracy, despite being somewhat less accurate on the training set (88% accuracy). This implies that Multilingual BERT generalizes more effectively to data that has not yet been seen, probably because it manages the various linguistic contexts found in the reviews more skillfully. The merits and trade-offs between the two models in the context of sentiment analysis across multilingual datasets are highlighted in this graph[20].

4.9 User Authentication

The user authentication system play a vital role in providing safe and personalized access to the sentiment analysis platform. The system must balance security and ease of use given the importance of customer reviews for a service such as Daraz. Therefore, it is essential to implement robust Sign In and Sign Up functionality. The Sign In process is intended for existing users to log in using their registered email and password. In this process, multiple validation layers are applied to ensure that the credentials provided by the user match the records stored in the database. If there is a mismatch, users are provided with an error message that provides guidance on how to rectify the error. In addition to password validation, the system utilizes advanced encryption techniques to ensure the security of sensitive data and passwords in order to minimize the risk of unauthorized access. Secure communication protocols such as HTTPS, which safeguard user data during transmission, enhance security. The Sign Up process is tailored to new users in order to enable them to create an account on the platform. It requires users to provide their full name, a unique email address, and a secure password. To ensure a high level of security, the password must meet specific complexity criteria, such as including uppercase and lowercase letters, numbers, and special characters. Once users submit the Sign Up form, the system verifies the uniqueness of the email and the complexity of the password. If all validations are met, the system creates a new user record and, in some cases, sends a confirmation email for further verification. This confirmation step is crucial to prevent spam and unauthorized account creation. The design of the user authentication system emphasizes both security and user experience. The user interface makes use of modern web technologies like HTML, CSS, Bootstrap, and JavaScript to provide a comfortable, responsive, and visually appealing experience. As a result of this design approach, the system can be used across a variety of devices, including desktop computers, tablets, and smartphones.

4.10 Technology Stack

The technology stack used in the development of the user authentication system consists of several key components that together provide a simple yet robust foundation. These technologies are commonly used in modern web development, ensuring compatibility and ease of maintenance. Here's an overview of each component and why it was chosen for this project:

- **HTML**: HyperText Markup Language (HTML) is the standard language for creating web pages. It serves as the structural backbone of the user interface, defining the elements that make up the Sign In and Sign Up pages. HTML allows for easy organization of page components such as forms, buttons, and links, providing a clear layout for users to interact with.

Figure 4.5: Html logo

- **CSS**: Cascading Style Sheets (CSS) is used to style the web pages, offering a way to control the visual appearance of HTML elements. CSS allows for customization of colors, fonts, spacing, and layout, contributing to a visually appealing user interface. By separating style from structure, CSS makes it easier to maintain and update the look and feel of the application.

Figure 4.6: Css logo

- **Bootstrap**: Bootstrap is a popular front-end framework that simplifies the development of responsive web pages. It provides pre-built components and a flexible grid system, making it easy to create interfaces that adapt to different screen sizes and devices. This is particularly important for a user authentication system, as users may access the platform from various devices such as desktops, tablets, or smartphones. Bootstrap's responsive design ensures a consistent user experience across all platforms.

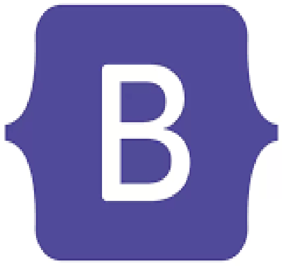


Figure 4.7: Bootstrap logo

- **JavaScript**: JavaScript is a versatile scripting language that adds interactivity and dynamic behavior to web pages. In the user authentication system, JavaScript is used to implement client-side validation, form submission, and user feedback. It allows for real-time validation of user inputs, providing immediate feedback on errors or incomplete information, thereby improving the user experience.

Figure 4.8: Javascript logo

4.11 Sign In Functionality

Existing users may authenticate themselves through the Sign In functionality to access the sentiment analysis platform. In order to ensure a seamless and secure login process, a series of intuitive steps is included. Here's a detailed breakdown of each component:

- **Email Field**: A user will be prompted to enter their registered email address in this essential component. The requirement that users provide their email addresses ensures that only authorized users with valid credentials are permitted access. Moreover, the email field includes a basic validation check to ensure that the format of the email address is correct, preventing errors that are common when entering an email address.

- **Password Field**: The password field is a critical aspect of the Sign In process, demanding users to input their designated password for authentication. Passwords serve as the primary means of verifying user identity, and as such, it's paramount that they remain confidential and well-protected. To safeguard user accounts against unauthorized access, the password field employs stringent security measures, including encryption techniques to obscure passwords from prying eyes. Additionally, the field may enforce password complexity requirements, such as minimum length and the inclusion of alphanumeric characters, to bolster security further.

- **Submit Button**: Authentication is initiated by clicking the submit button. Users can proceed with the login attempt once they enter their email address and password. Credentials are verified against user data stored in the database behind the scenes. The sentiment analysis platform is accessible to users if their credentials match, allowing them to explore and utilize its features. A prompt error message prompts users to correct their input if the authentication fails due to incorrect credentials.

We prioritize simplicity, efficiency, and security when designing the Sign In functionality. A streamline login process along with robust security measures ensures that users can register quickly and confidently, knowing that their personal information remains secure. Additionally, the Sign In functionality sets the stage for a seamless user experience, laying the groundwork for users to engage with the sentiment analysis platform effortlessly.

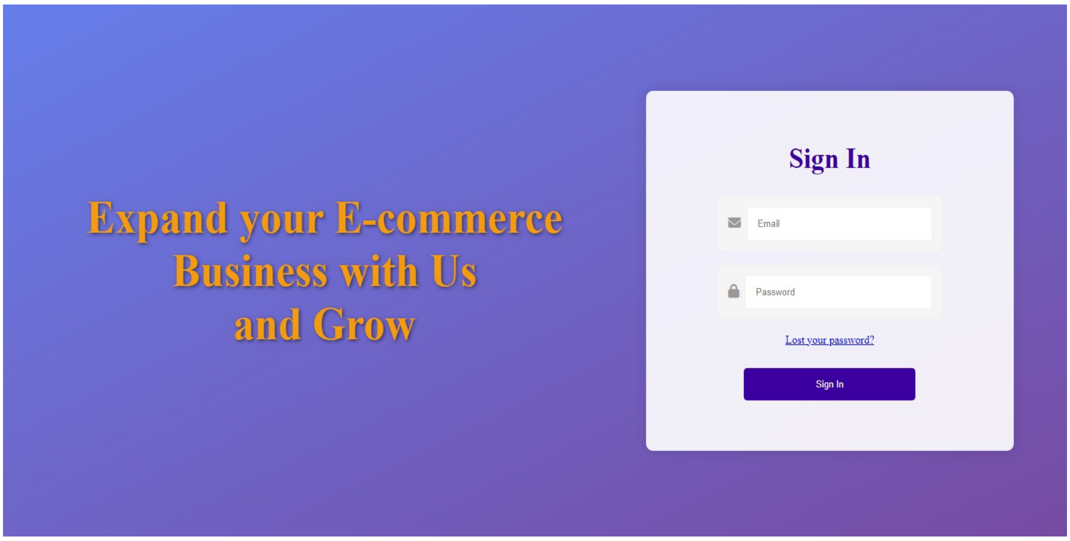


Figure 4.9: SignIn Page

4.12 Sign In Form Validation

The validation of forms is an essential component of the Sign In functionality in order to ensure data integrity and user experience. Before submitting entries for authentication, this tool verifies that user inputs, particularly email addresses and passwords, are accurate and complete. Let's delve deeper into how this process unfolds:

- **Validation Criteria**: There are a number of predefined criteria that determine what constitutes a valid input for each field during the form validation process. It is typically required to check for the presence of the "@" symbol in the email address field, the presence of a domain name following the "@" symbol, and the absence of any spaces in the address.
- **Client side validation**: When JavaScript functions are used to evaluate the validity of a user's input in time, when it is interacting with Sign In forms, form validation starts on the client side. To determine whether the entered email address and password meet the specified validation criteria, these JavaScript functions will perform a series of checks. For example, JavaScript will immediately detect these errors and trigger corresponding error messages when the user attempts to submit a form with an empty email address field or password that does not comply with complex requirements. These error messages are displayed, providing users with immediate feedback on their inputs and guidance to correct them, in an inline format next to the corresponding input fields.
- **Reduced Server-side Processing**: By conducting validation checks on the client side, the form validation process helps alleviate the workload on the server. Before transmitting form data to the server, it is possible to identify and address a number of common validation errors such as empty fields or incorrect formatting inputs on the client side.
- **Increased user experience**: The overall user experience is greatly enhanced by

immediate feedback from the client side as a result of validation. When users are interacting with the Sign In form, they will receive immediate validation results that enable them to immediately detect and correct any errors on their input.

4.13 Authentication Process

After filling in the email and password fields in the Sign In form, the authentication process is started to check their identity and allow the user to access our main page of the website where he can get a sentimental analysis of the product.

4.14 Data Transmission

After pressing the Sign In button, the user's email and password are transmitted safely over a secure communication protocol, such as HTTPS, to encrypt the data and prevent interception by unauthorized parties. The System safeguards user credentials against potential security threats and ensures the confidentiality of sensitive information.

4.15 Server-side Verification

Upon receiving the user's data, the server starts the verification process by comparing the user's submitted email address and password with the stored records in the database. This process involves retrieving the user's information from the database through a query.

4.16 Credential Comparison

Once the data is received, the server checks the both email and password are the same or not, if both are the same, the user will log in otherwise he will be redirected to the same page again. This comparison is performed using a hashing algorithm such as bcrypt.

4.17 Sign Up Functionality

The Sign Up functionality basically act as the entry point for new users to join our sentiment analysis platform. This functionality enables them to create their account free of cost and gain access to our website features. This process is created so that we have data of our users. We can give them suggestions that which product is trending now. This will help the users to buy products without wasting their important time.

- **Name Field**: In this field, user will provide the full name during the account creation process. The Name field regularly checks to ensure that users enter a valid name format and prevent any potential input errors.

- **Email Field**: The email field hold users to enter a valid email address, which act as their unique identifier within the platform. This field employ email validation techniques, such as checking for the presence of the "@" symbol and a valid domain name, to ensure the accuracy of user provided email addresses.

- **Password Field**: The Password field allow users to set a secure password for their account, safeguarding their personal information and ensuring the integrity of their account credentials. Weak or guessable passwords may rise a risk to your user account. To keep your account secure, keep your password should be comprise of lowercase letters, uppercase letters, numbers and special characters.

- **Submit Button**: This button is the final step towards user's account creation. Upon clicking this button, the platform checks the user's inputs and after ensuring that all fields meet the criteria. Once the checking complete, the platform creates the new user account in its database, assigning a unique identifier.

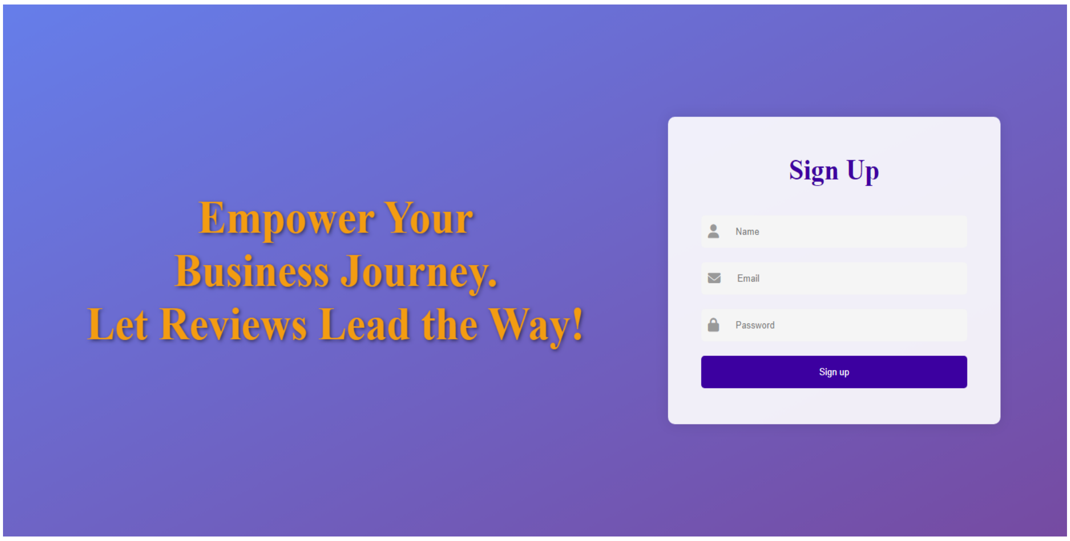


Figure 4.10: SignUp Page

4.18 Sign Up Form Validation

In Sign Up, form validation plays a vital role. This servers as a first line of defense against the malicious or incorrect data. By ensuring strict validation checks, this ensures the accuracy, integrity and security of user provided data.

- **Name Validation**: The Name field validation ensures that it should not be empty. This is because the user's name is essential for personalized interactions and identification within the platform.
- **Email Validation**: The Email Validation process ensures the presence of "@" symbol and a valid domain name.
- **Password Validation**: The Password field play role in securing the user's account and save the sensitive information within the platform. To keep information save, your password should be the combination of uppercase letters, lowercase letters, numbers and special characters.

4.19 Account Creation Process

Upon clicking the Sign Up button, system checks that email address should contain "@" symbol and a valid domain name and password should be unique. Validation passes and a new user record should be created in the database.

4.20 User Experience and Design Principles

We kept the design of Sign Up and Sign In form simple and easy to fill. The key principles that we follow to design these forms are as follows:

- **Simplicity**: We follow the simplest layout, minimizing complexity and confusion.
- **Responsive**: Our both forms are responsive, they can adjust themselves accordingly to their screen sizes, providing a consistent experience across devices.

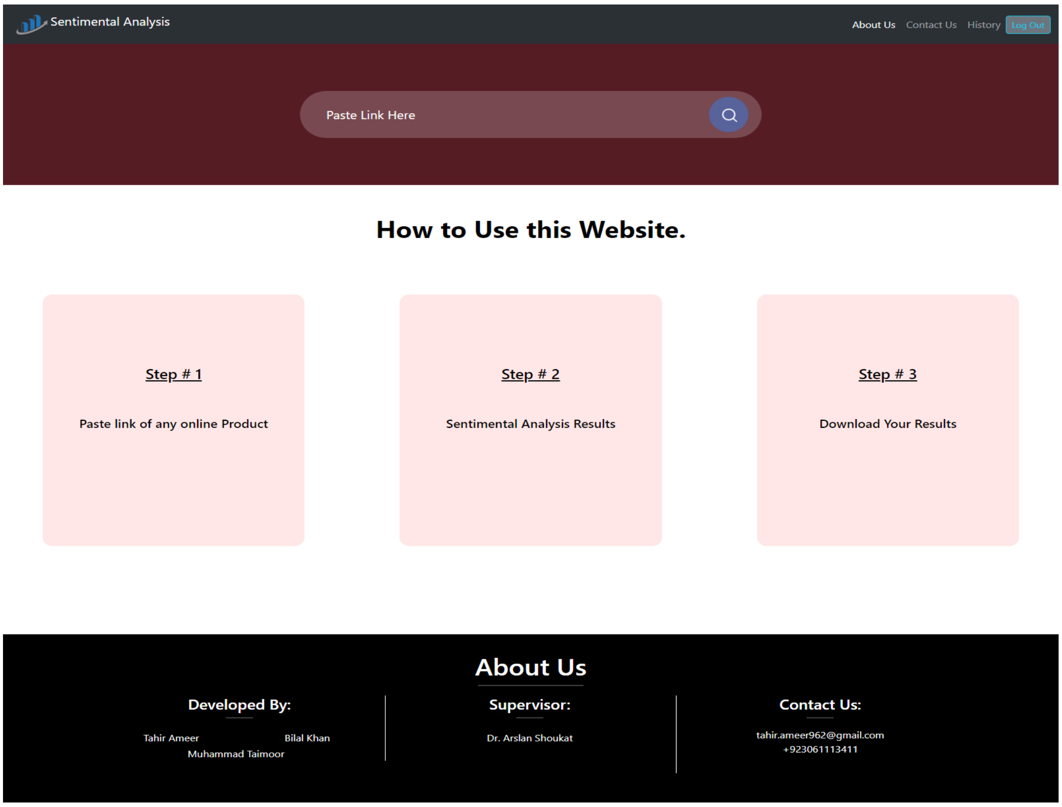
4.21 Future Improvements

- **Two-Factor Authentication (2FA)**: We are planning to add an additional layer of security by requiring the user's to provide a two factor authentication, such as a 4 digit number sent to their mobile device.
- **Social Media Integration**: We are also planning to facilitate the user's to sign in through their social media accounts for a seamless experience.

These improvements would greatly enhance the security and user experience of the user. This would contribute to a more robust and reliable platform.

4.22 Sentiment Analysis Platform

After logging into the account, the user's will see the main page where he/she can get the sentiment analysis of the desired product. We have mentioned the steps on our website on how to use it. User's will simply paste the link of any online daraz's product of which he wants to know whether he should buy this product or not. After pressing the submit button, our website will approach that product link and starts collecting all the reviews of the previous customers in an excel file through web scraping. After that, our NLP model Distil Bert Starts its working and classify the reviews into three categories Positive, Negative and Neutral. Once we get the percentages of reviews, we will display that into two types of graphs Pie chart and Bar chart.



How to Use this Website.

Step # 1

Paste link of any online Product

Step # 2

Sentimental Analysis Results

Step # 3

Download Your Results

About Us

Developed By:

Tahir Ameer Bilal Khan
Muhammad Taimoor

Supervisor:

Dr. Arslan Shoukat

Contact Us:

tahir.ameer962@gmail.com
+923061113411

Figure 4.11: Main Page

- **Bar Chart**: Bar Charts is mostly used for easy comparisons between different categories. It gives us a clear vision. You do not need specialized knowledge to clarify a bar chart. It is completely understandable across different cultures and disciplines.

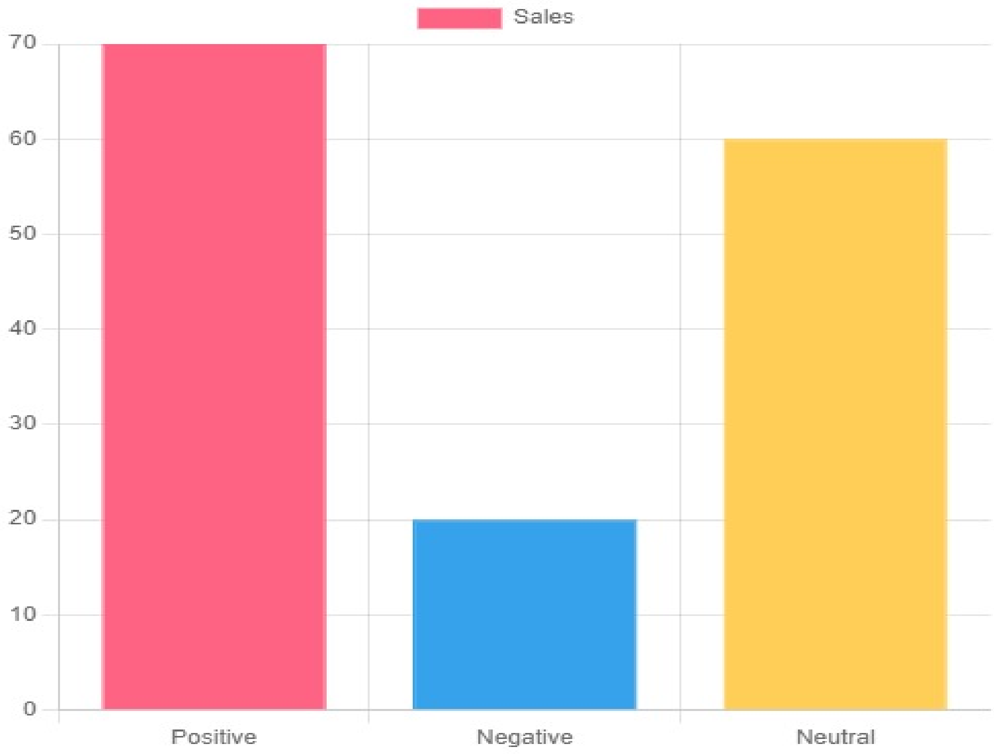


Figure 4.12: Bar Chart

- **Pie Chart**: Pie Chart represents the information by dividing a circle into slices, with each slice representing a portion of the whole, In our case there will be three portions, Positive, Negative and Neutral. Pie charts are simple and organized which makes them easy to interpret immediately.

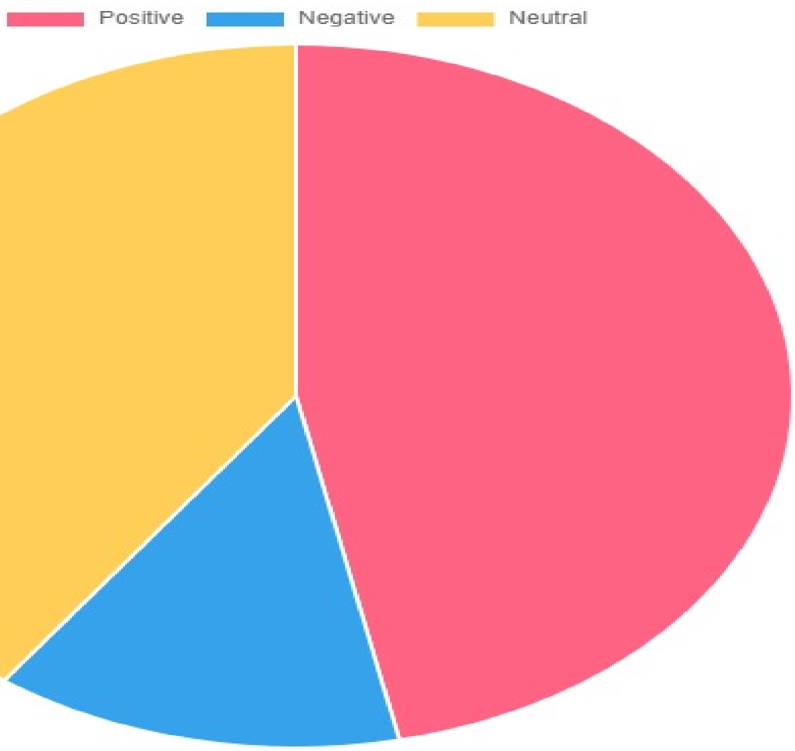


Figure 4.13: Pie Chart

4.23 Web Scraping

Web Scraping is one of the main tool in our project, as it enable us to collect reviews from the Daraz website for the sentimental analysis of the reviews. Selenium, a powerful tool is used for automating web browsers.

4.23.1 Scraping Techniques

4.23.1.1 Tools and Libraries Used

- **Selenium**: Selenium is a tool which is used for automating web browsers. Its make it possible to navigate through web pages and extract the necessary data from the pages.
- **Selenium WebDriver**: Selenium WebDriver is used for controlling web browsers

programmatically. It plays a vital role in functioning tasks like clicking buttons, filling out forms, and scrolling through pages.

- **Python**: Python is a programming language which is used for writing scripts for different purposes like for image processing, signal processing, and for many other fields. It also plays a role in writing the scripts that control Selenium.
- **BeautifulSoup**: BeautifulSoup can be used for parsing the HTML and pull out the data effectively when Selenium is used for routing and interaction.

4.24 Contact Us

Contact us page is a main component of the sentimental analysis platform. This page is designed to help communication between users and us. This act as a mode of communication and to give feedback about our sentimental analysis platform.

4.24.1 Contact Us Form

Main feature of "Contact Us" page is the contact us form. In this form we collect the essential information of the user in a descent manner. This form typically contains:

- **Name**: A field for users to enter its name. This allows the support team to address users problem personally.
- **Email**: A mandatory field for users to enter their email address. This is crucial for the support team to respond to user questions through this email.
- **Query Box**: This message box is a vital component of the contact form, providing users with a flexible and comprehensive space to describe their issues, ask questions, or give feedback.

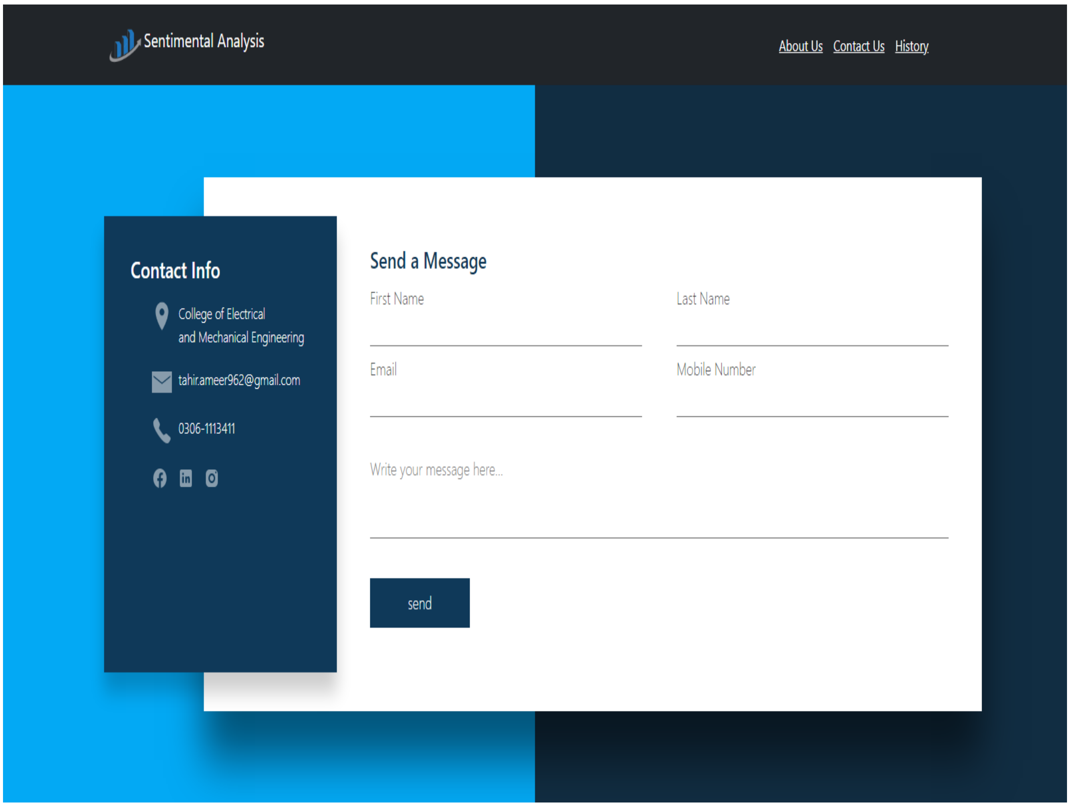
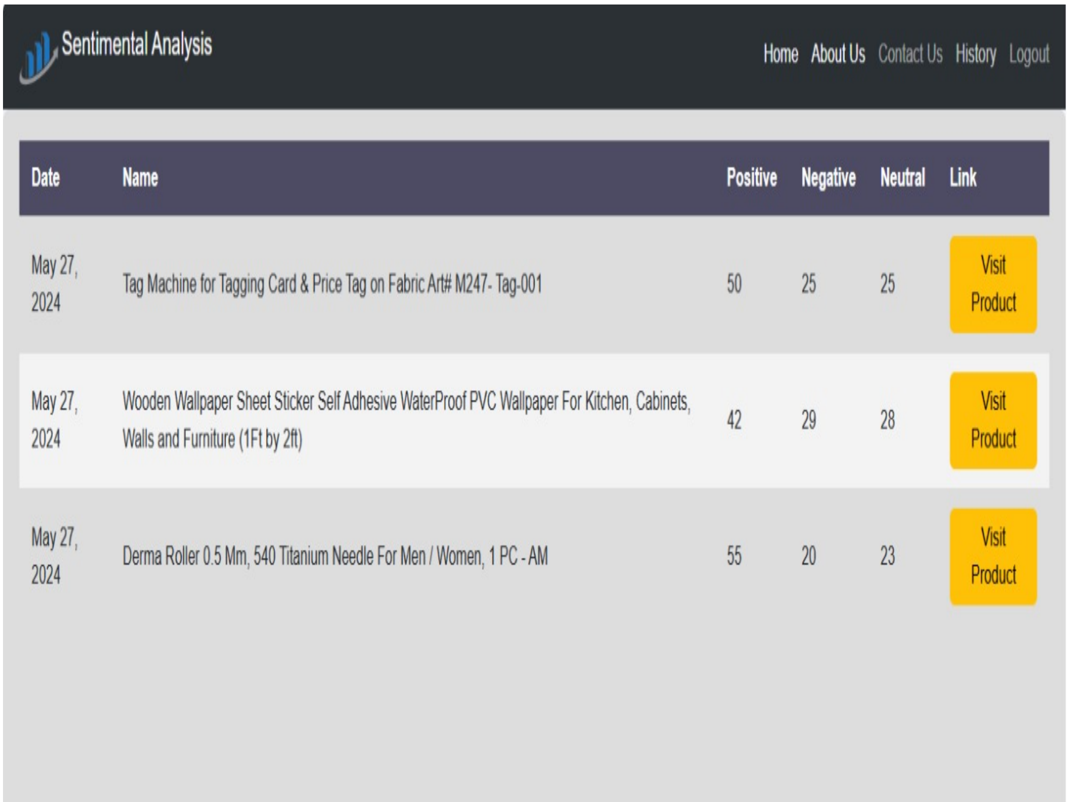


Figure 4.14: Main Page

4.25 History Page

Here we store the data of the products that users have searched throughout their browsing history. There are five columns in this form that contain the names of serial numbers. There is a link to the product, the name of the product, positive and negative reviews, neutral reviews, and the price of the product. Our goal is to keep a record of the products that have been searched most frequently in recent times by using this page. We will be able to assist the sellers by informing them that the product mentioned in this article is in high demand at the moment. Therefore, you should keep this product in your store as much as possible. Similarly, through this, we will also help the users to give suggestions that this product is the most sold in this month. And the previous customer's experience was also good about this product. So the history page is playing the most important role for us to help both customers and sellers at the same time.



Date	Name	Positive	Negative	Neutral	Link
May 27, 2024	Tag Machine for Tagging Card & Price Tag on Fabric Art# M247- Tag-001	50	25	25	Visit Product
May 27, 2024	Wooden Wallpaper Sheet Sticker Self Adhesive WaterProof PVC Wallpaper For Kitchen, Cabinets, Walls and Furniture (1Ft by 2ft)	42	29	28	Visit Product
May 27, 2024	Derma Roller 0.5 Mm, 540 Titanium Needle For Men / Women, 1 PC - AM	55	20	23	Visit Product

Figure 4.15: History

4.26 Database Design and Implementation

In the implementation of sentimental analysis platform for reviews on daraz, a well-established database is required. This part outlines the design and implementation of the database that we are using in the project, highlighting the structure of the user and product tables. We are using the inbuilt SQLite database system in Django, chosen for its simplicity and ease of integration with the Django framework.

4.27 Database Selection

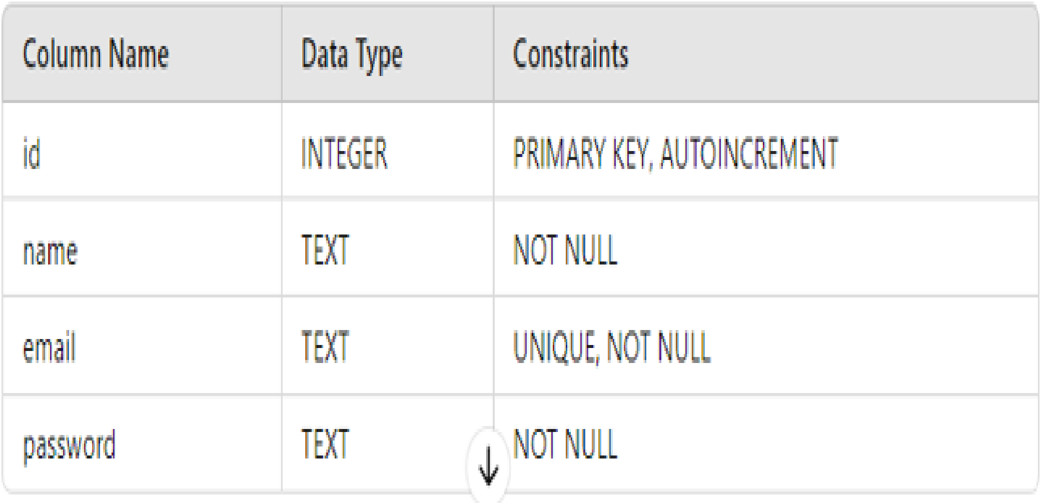
We selected SQLite database because we are already using the Django framework for our website. Secondly best part of this website is its lightweight nature and seamless integration with Django. As this database does not require any separate server process, which makes very easy development and deployment, particularly for our website.

4.28 Database Structure

Our database is comprised of two primary tables “User” and “Product”. These tables stores the information of the related user account and product and its reviews.

4.28.1 User Table

The “User” table is designed to keep the information about the users of our website. Following will the structure of our “User” table:



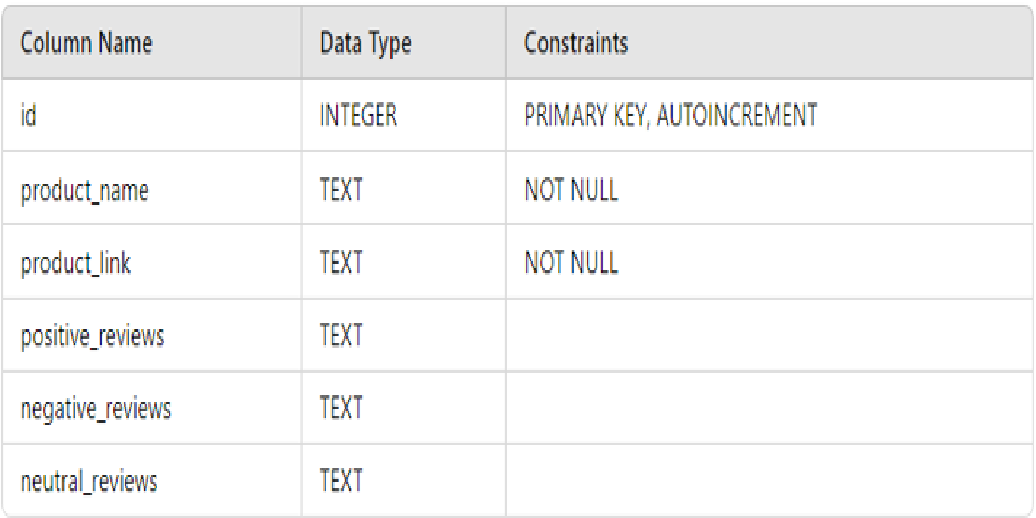
Column Name	Data Type	Constraints
id	INTEGER	PRIMARY KEY, AUTOINCREMENT
name	TEXT	NOT NULL
email	TEXT	UNIQUE, NOT NULL
password	TEXT	NOT NULL

Figure 4.16: User Table

- **ID**: This is a primary key for user table. A unique number for each user. This number will be generated automatically.
- **Name**: This part stores the name of the user.
- **Email**: The email address of the user , this part servers as the unique identifier for the login purposes.
- **Password**: This field stores the password of the user to ensure security and privacy.

4.28.2 Product Table

The “Product” table stores the information of the product. The structure of the product table is as follows:



Column Name	Data Type	Constraints
id	INTEGER	PRIMARY KEY, AUTOINCREMENT
product_name	TEXT	NOT NULL
product_link	TEXT	NOT NULL
positive_reviews	TEXT	
negative_reviews	TEXT	
neutral_reviews	TEXT	

Figure 4.17: Product Table

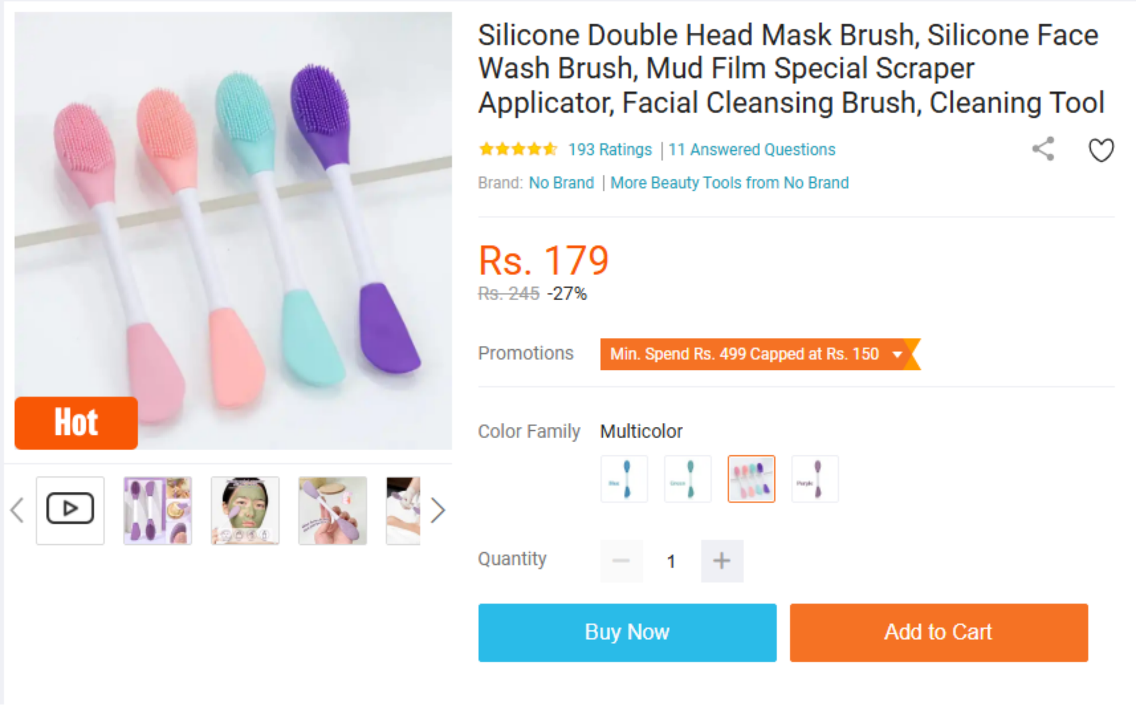
- **ID**: This is a primary key for the product table. A unique number for each product. This number will be generated automatically.
- **Product Name**: The name of the product..
- **Product Link**: The URL link to the product page on Daraz.
- **Positive Reviews**: This part of the table contains the positive reviews for the product.
- **Negative Reviews**: A part of the table contains negative reviews for the product.
- **Neutral Reviews**: A part of the table contains neutral reviews for the product.

4.29 Results Comparison with Daraz Star Ratings

The main goal of this study is to assess this reviews and compare them with customers ratings. The research will seek to determine whether there is a divergence between text based sentiment and numerical ratings, and the implications for the both the credibility and trustworthiness of online reviews.

4.29.1 Silicon Double Head Mask Brush

This is a product on daraz, lets compare this product's star rating with our website's result.



Silicone Double Head Mask Brush, Silicone Face Wash Brush, Mud Film Special Scraper Applicator, Facial Cleansing Brush, Cleaning Tool

★★★★★ 193 Ratings | 11 Answered Questions

Brand: No Brand | More Beauty Tools from No Brand

Rs. 179

~~Rs. 245~~ -27%

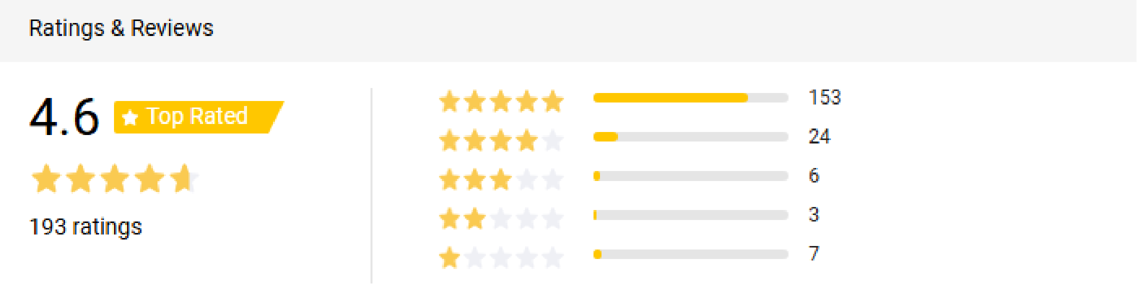
Promotions: Min. Spend Rs. 499 Capped at Rs. 150

Color Family: Multicolor

Quantity: 1

Figure 4.18: Daraz Product

Star rating of this product is **4.6** on the website.



Ratings & Reviews

4.6 ★ Top Rated

★★★★★

193 ratings

★★★★★	153
★★★★☆	24
★★★☆☆	6
★★☆☆☆	3
★☆☆☆☆	7

Figure 4.19: Star Rating

Now, Let's compare this star rating with our website results.

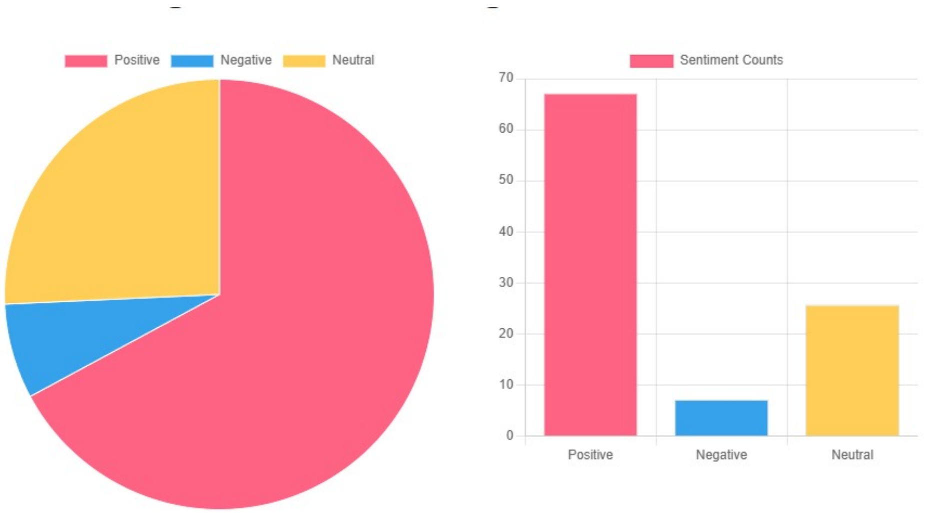
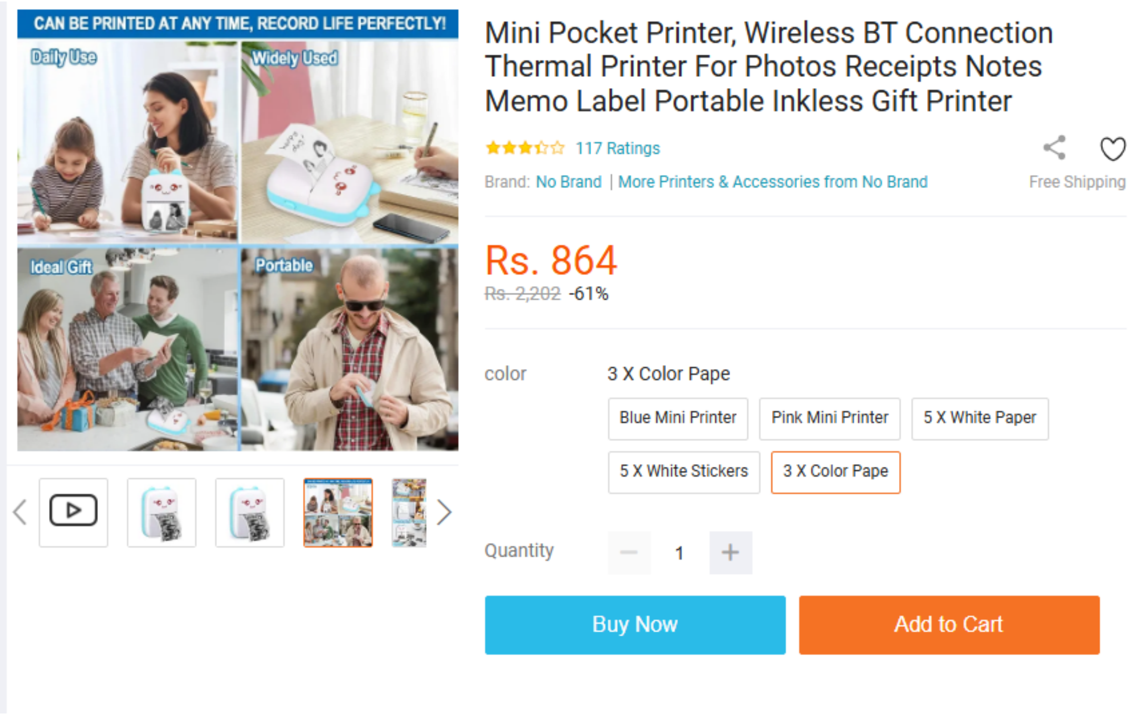


Figure 4.20: Sentiments Analysis Platform Result

This clearly shows that our Sentiments Analysis Platform predict the results that matches with the Daraz star rating.

4.29.2 Mini Pocket Printer

This is a product on daraz, lets compare this product's star rating with our website's result.



Mini Pocket Printer, Wireless BT Connection Thermal Printer For Photos Receipts Notes Memo Label Portable Inkless Gift Printer

117 Ratings

Brand: No Brand | More Printers & Accessories from No Brand

Free Shipping

Rs. 864

~~Rs. 2,202~~ -61%

color 3 X Color Pape

Blue Mini Printer | Pink Mini Printer | 5 X White Paper | 5 X White Stickers | 3 X Color Pape

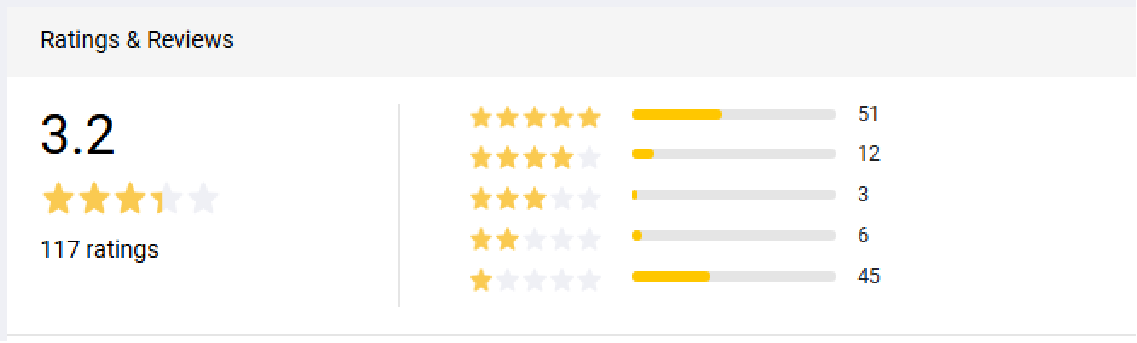
Quantity − 1 +

Buy Now

Add to Cart

Figure 4.21: Daraz Product

Star rating of this product is **3.2** on the website.



Ratings & Reviews

3.2

117 ratings

★★★★★	51
★★★★☆	12
★★★☆☆	3
★★☆☆☆	6
★☆☆☆☆	45

Figure 4.22: Star Rating

Now, Let’s compare this star rating with our website results.

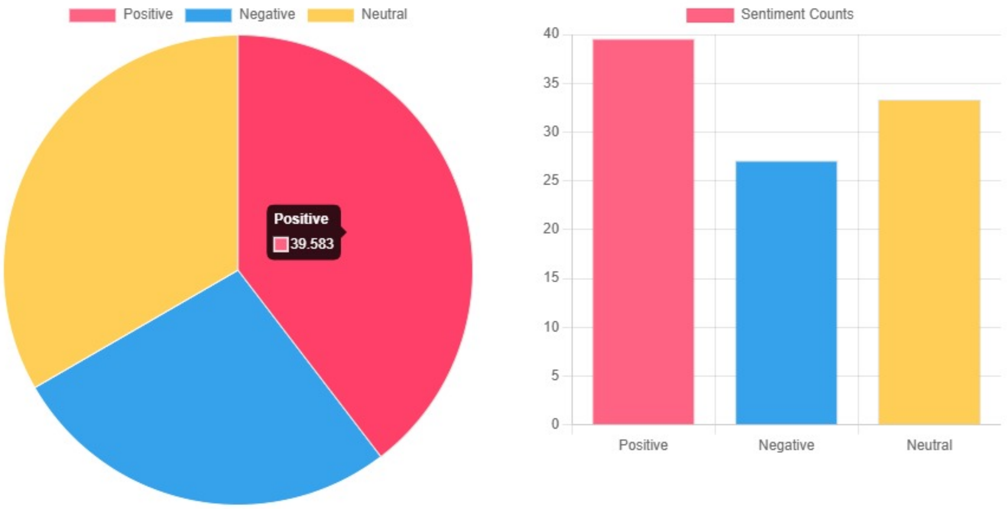


Figure 4.23: Sentiments Analysis Platform Result

We have performed sentiment analysis and found that it correlates well with the star rating; so a sentiment extracted from the review corresponds to the numerical rating given. Such congruence indicates fairly that the rating system of the platform is objective and can be taken at face value. So for customers and brands alike, they can be sure that the numbers and feedback they see on Daraz are the real, real deal. Daraz continues to be a trustworthy outlet for real customer reviews, and this uniformity enforces its credibility.

Chapter 5

Market Analysis

Commercializing a project requires performing detailed market research to create a sustainable business plan for it. Market analysis helps in figuring out the need as well as the demand for the product. The key objective is to determine that there indeed is a market where this product can be sold and will have high demand. This will enable us to add more features and enhance the product further, thus encouraging further research and development.

5.1 Market Size

The market size is considerable as there are a lot of people in Pakistan who take part in online shopping, and this number is only projected to grow in the coming years. Statistics from Daraz.pk tell us that by 2030, Pakistan will become the 7th largest consumer market in the world, and is a country with internet penetration growing 20% each year[**?**]. This projected increase in the trend of online shopping will only increase with time in Pakistan. As such, our product has a huge market in the country.

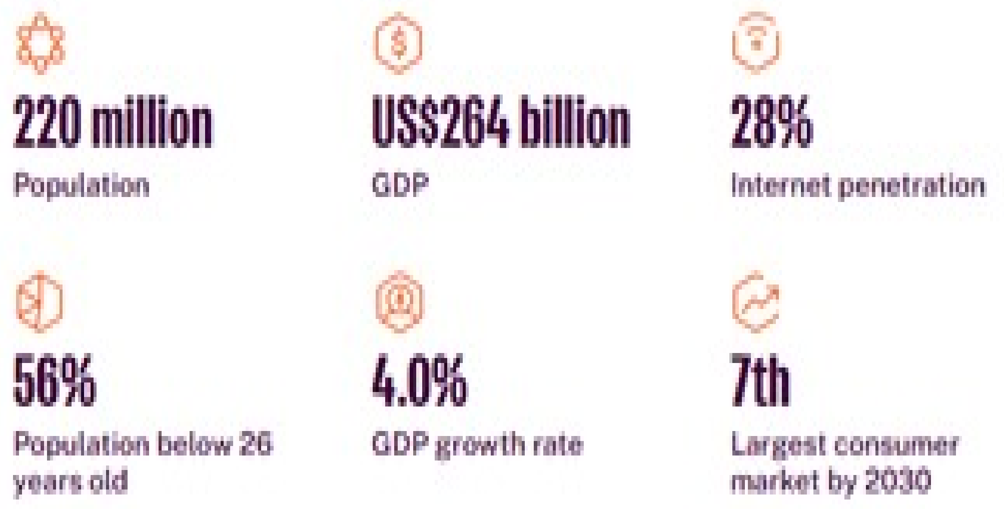
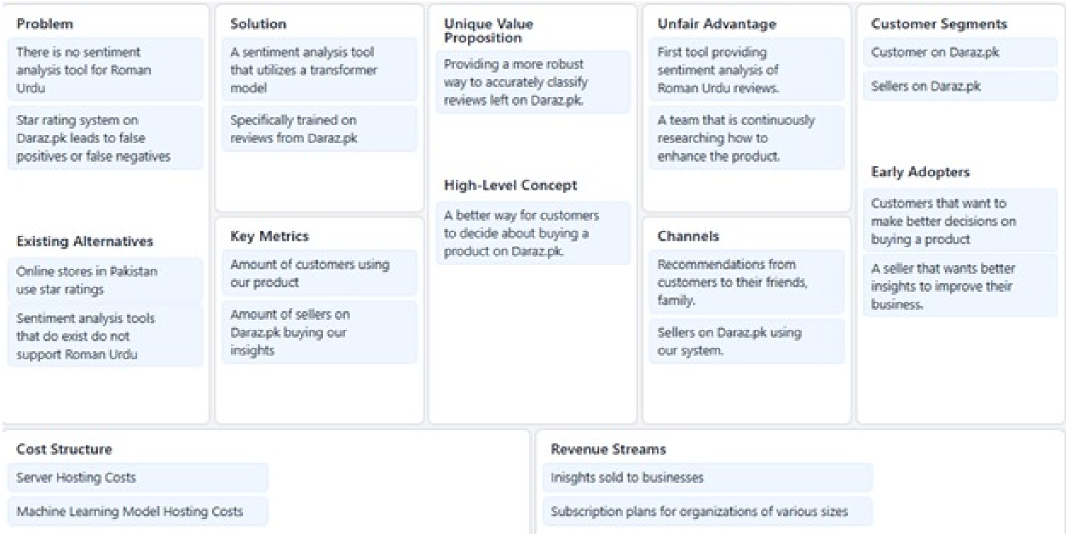


Figure 5.1: The projected trend of increase in online shopping

5.2 Business Model

For a newly created company, a business model is critical. A business model ensures that a company will thrive after launching something in the market. Creating a business model for a startup or product means identifying the problem it is going to solve, the market that will be served, the level of investment required, what products or services will be offered, and how will the revenue be generated. A business plan is a lengthy document, and as such, it will not be possible to accommodate it in this report. As an alternative, we can make a lean canvas model of our business. This will help in highlighting the most significant aspects of our business within one page. The lean canvas methodology gives a one-page business plan that lets investors have a quick overview of the entire working of the business and its potential outcomes. The Lean Canvas Model of our project is shown below:



Problem

There is no sentiment analysis tool for Roman Urdu

Star rating system on Daraz.pk leads to false positives or false negatives

Existing Alternatives

Online stores in Pakistan use star ratings

Sentiment analysis tools that do exist do not support Roman Urdu

Solution

A sentiment analysis tool that utilizes a transformer model

Specifically trained on reviews from Daraz.pk

Key Metrics

Amount of customers using our product

Amount of sellers on Daraz.pk buying our insights

Unique Value Proposition

Providing a more robust way to accurately classify reviews left on Daraz.pk.

High-Level Concept

A better way for customers to decide about buying a product on Daraz.pk.

Unfair Advantage

First tool providing sentiment analysis of Roman Urdu reviews.

A team that is continuously researching how to enhance the product.

Channels

Recommendations from customers to their friends, family.

Sellers on Daraz.pk using our system.

Customer Segments

Customer on Daraz.pk

Sellers on Daraz.pk

Early Adopters

Customers that want to make better decisions on buying a product

A seller that wants better insights to improve their business.

Cost Structure

Server Hosting Costs

Machine Learning Model Hosting Costs

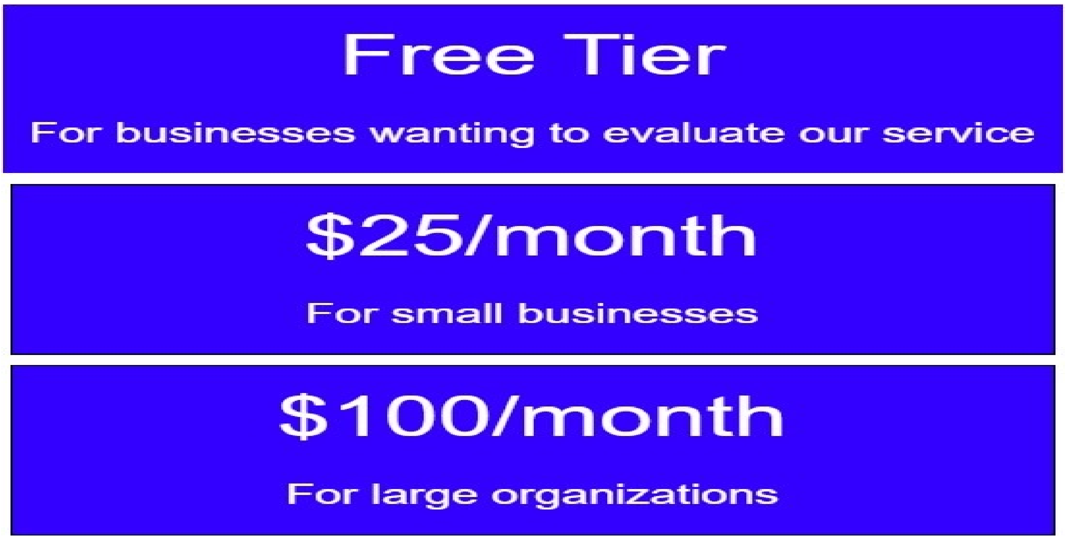
Revenue Streams

Inisghts sold to businesses

Subscription plans for organizations of various sizes

Figure 5.2: Lean Canvas Model of our sentiment analysis tool for Daraz.pk

It can be seen from Table 1 that market prices of competing products are quite high compared to the tool we developed. Moreover, none of the tools listed in Table 1 explicitly provide support for Roman Urdu. To help small businesses we plan on giving a free trial of one month before we charge them for our services. The enterprise version of our product, however, will have more features that are suited for processing data of Roman Urdu text at much larger volumes.



Free Tier

For businesses wanting to evaluate our service

$25/month

For small businesses

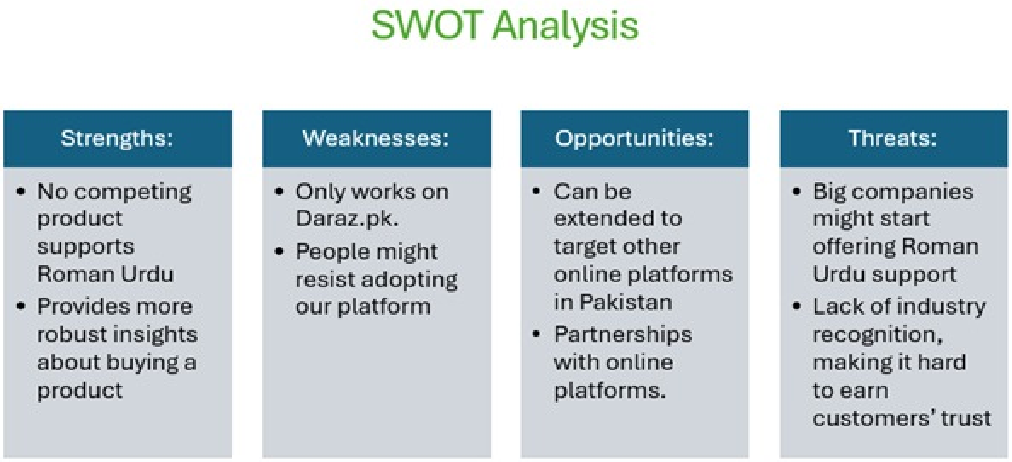
$100/month

For large organizations

Figure 5.3: Lean Canvas Model of our sentiment analysis tool for Daraz.pk

5.3 SWOT Analysis

A SWOT analysis is helpful for a business to analyze its strengths, weaknesses, opportunities and threats. Having an idea of such things helps businesses to plan accordingly to be successful in the market. Our SWOT analysis shows that although we have existing threats such as sentiment analysis tools created by big companies like IBM and Amazon and weaknesses but the opportunity at hand is too good to be thrown away. A market that has not been explored at a large scale, as well as the existence of online platforms that have reviews in the form of Roman Urdu makes it a very good opportunity to establish a profitable business.



SWOT Analysis

Strengths:

- No competing product supports Roman Urdu
- Provides more robust insights about buying a product

Weaknesses:

- Only works on Daraz.pk.
- People might resist adopting our platform

Opportunities:

- Can be extended to target other online platforms in Pakistan
- Partnerships with online platforms.

Threats:

- Big companies might start offering Roman Urdu support
- Lack of industry recognition, making it hard to earn customers' trust

Figure 5.4: SWOT analysis

5.4 Innovative Comparison

The innovation in our product cannot be debated. On a scale where we put affordability against Roman Urdu support, this is an innovative advancement among the many sentiment analysis tools in the market. There is a very huge gap that needs to be filled, which means that we have everything needed to turn our product into a successful business.

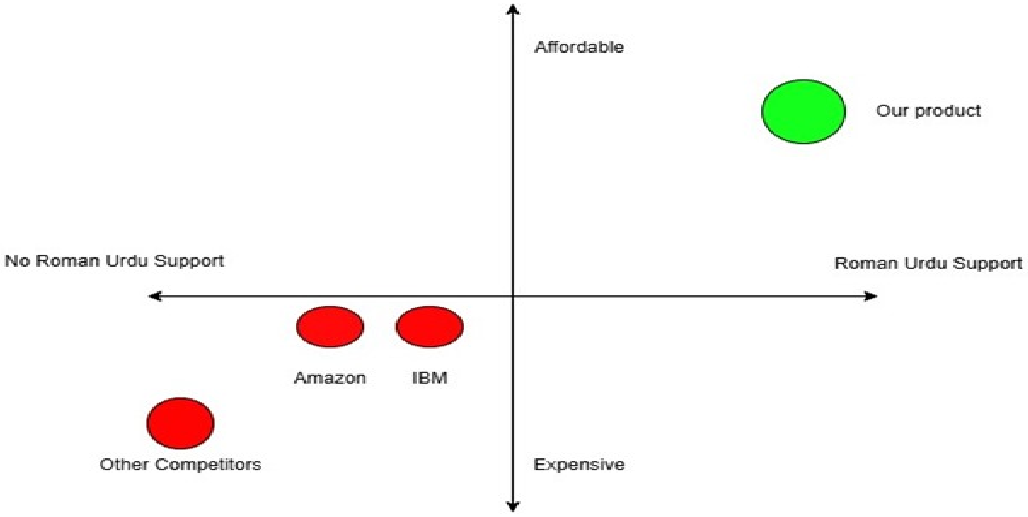


Figure 5.5: Innovative comparison with other sentiment analysis tools

Chapter 6

Conclusion

Analyzing client sentiment on e-commerce platforms such as Daraz is crucial in gaining insights into customer preferences, satisfaction levels, and areas for potential improvement. This project aimed to develop a sentiment analysis tool specifically for Daraz using state-of-the-art machine learning models, sophisticated web scraping methods, and efficient database administration processes. The objective is to develop a strong framework that could reliably classify customer reviews into positive, negative, and neutral attitudes, offering insightful information to different Daraz ecosystem players.

6.1 Methodology and Model Refinement

The foundation of this sentiment analysis tool is built upon fine-tuning pre-trained language models, including DistilBERT and Multilingual BERT, using a specially gathered and tagged dataset from Daraz. This phase is crucial in customizing the models to understand the unique complexities of language used in Daraz product reviews. These models improve the ability to precisely represent and decipher the emotions that customers convey in their evaluations. DistilBERT's pre-trained capabilities provides a solid base for sentiment analysis. However, fine-tuning the model on a dataset specific to Daraz reviews which mostly consist of Roman Urdu significantly improves its accuracy and relevance.

The model's ability to comprehend the subtle contextual details present in Daraz evaluations enables it to classify attitudes more accurately.

6.2 Real-Time Data Collection and Analysis

To ensure our sentiment analysis is grounded in real-time data, it has integrated a sophisticated web scraper capable of automatically collecting reviews whenever customers input a product link from Daraz. This real-time data acquisition strategy is essential in maintaining the relevance and timeliness of our sentiment analysis. Dynamic pipelines are established using internet that allows us to collect data continuously and feed it straight into our sentiment analysis model. Reviews are processed quickly and easily into positive, negative, and neutral sentiment categories because of the DistilBERT model's API interface. This real-time analysis capability enables us to capture and analyses customer feedback as it evolves, providing a dynamic and up-to-date view of customer sentiment on the Daraz platform.

6.3 Benefits and Implications

The deployment of our sentiment analysis tool brought several significant benefits to various stakeholders within the Daraz ecosystem:

- **Accuracy and Customizability**: Accuracy: DistilBERT's accuracy in sentiment categorization improves optimization of the model with Daraz-specific data. This increase in precision makes sure that the feelings found in the evaluations are almost the same as the real feelings that the clients has stated.
- **Customizability**: The use of a web scraper (selenium) to analyze reviews based on specific product links allows the users to obtain targeted insights. This capability facilitates a more granular understanding of customer sentiments toward a particular product, aiding in more informed decision-making.

- **Scalability and Integration**: It is feasible to ensure quick and easy integration with a variety of applications by using DistilBERT's API. The sentiment analysis tool's scalability allows additional features and future improvements.
- **Data Collection and Storage**: A well-built data collection pipeline guarantees a steady flow of important data for analysis. Deeper insights into changing customer sentiment over time is made possible by longitudinal research and trend analysis is possible by the systematic database storage of this data.

6.4 Impact on Stakeholders

The sentiment analysis tool affects a wide range of stakeholders, including Daraz platform users, sellers, and customers:

- **Consumers**:Customers use the tool to learn more about the sentiment surrounding a product before making a purchase. The platform is more trustworthy and confident by the openness and accessibility of sentiment data, which improves the entire user experience.
- **Sellers**:The tool is used by sellers to gauge consumer satisfaction levels with their products. Sellers can resolve customer complaints, improve the caliber of their goods, and eventually improve their standing in the market by recognizing unfavorable comments.
- **Daraz Platform**:Daraz determines opportunities for platform enhancement and provide important insights into consumer satisfaction by examining general sentiment trends. Daraz uses this data to make strategic decisions that enhance user experience and platform growth.

6.5 Conclusion

This project shows how effective it is to combine sophisticated machine learning models, strong database administration, and real-time data collection to develop a potent sentiment analysis tool for the Daraz e-commerce platform. The website creates the opportunity to correctly categorize product reviews into positive, negative, and neutral attitudes by utilizing DistilBERT and multilingual fine-tuning. In the end, this feature of online purchasing is more successful and informs the offering insightful information to buyers, vendors, and the platform itself.

The achievement of high accuracy in sentiment categorization makes possible the fine-tuning of DistilBERT using data peculiar to Daraz. This modification ensures the model to understand the intricacies of language used in Daraz evaluations, which led to more accurate sentiment analysis. The sentiment analysis is based on the most recent data available, thanks to the integration of an advanced web scraper that makes the real-time data collection possible.

The research demonstrates how crucial it is to collect data in real-time and perform dynamic analysis to capture the dynamic nature of customer sentiments. The foundation led the sellers and buyers to provide more research and continuous model improvement by organizing the collected data into a well-organized database. This database proves to be an invaluable tool for carrying out longitudinal research, monitoring shifts in customer perception over time, and spotting new patterns in Daraz user activity.

References

[1] N. Azhar and S. Latif, “Roman urdu sentiment analysis using pre-trained distilbert and xlnet,” in *2022 Fifth International Conference of Women in Data Science at Prince Sultan University (WiDS PSU)*, pp. 75–78, 2022.

[2] M. Talat, H. Asim, and A. Asmat, “Classification of sentiments of the roman urdu reviews of daraz products using natural language processing approach,” in *2021 International Conference on Innovative Computing (ICIC)*, pp. 1–6, 2021.

[3] M. A. Qureshi, M. Asif, M. F. Hassan, A. Abid, A. Kamal, S. Safdar, and R. Akbar, “Sentiment analysis of reviews in natural language: Roman urdu as a case study,” *IEEE Access*, vol. 10, pp. 24945–24954, 2022.

[4] M. K. Nazir, M. Ahmad, H. Ahmad, M. Abdul Qayum, M. Shahid, and M. A. Habib, “Sentiment analysis of user reviews about hotel in roman urdu,” in *2020 14th International Conference on Open Source Systems and Technologies (ICOSST)*, pp. 1–5, 2020.

[5] A. Younas, R. Nasim, S. Ali, G. Wang, and F. Qi, “Sentiment analysis of code-mixed roman urdu-english social media text using deep learning approaches,” in *2020 IEEE 23rd International Conference on Computational Science and Engineering (CSE)*, pp. 66–71, 2020.

[6] Daraz, “Wave candy silicone phone case for iphone 11 12 13 14 pro max 7 8 plus

se 2020 xs max xs xr covers shockproof phone casing shell..” https://daraz.com/markets/pakistan/, 2024. Accessed: 2017-05-25.

[7] Amazon, “Amazon comprehend. retrieved from amazon web services.” https://aws.amazon.com/comprehend/.

[8] MonkeyLearn, “Monkeylearn.” https://monkeylearn.com/.

[9] IBM, “Ibm watson natural language processing, howpublished = https://www.ibm.com/products/natural-language-understanding,” 2017. Accessed: 2017-03-27.

[10] J. Singh, B. McCann, R. Socher, and C. Xiong, “BERT is not an interlingua and the bias of tokenization,” in *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)* (C. Cherry, G. Durrett, G. Foster, R. Haffari, S. Khadivi, N. Peng, X. Ren, and S. Swayamdipta, eds.), (Hong Kong, China), pp. 47–55, Association for Computational Linguistics, Nov. 2019.

[11] B. Büyüköz, A. Hürriyetoğlu, and A. Özgür, “Analyzing ELMo and DistilBERT on socio-political news classification,” in *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020* (A. Hürriyetoğlu, E. Yörük, V. Zavarella, and H. Tanev, eds.), (Marseille, France), pp. 9–18, European Language Resources Association (ELRA), May 2020.

[12] V. Dogra, A. Singh, S. Verma, Kavita, N. Z. Jhanjhi, and M. N. Talib, “Analyzing distilbert for sentiment classification of banking financial news,” in *Intelligent Computing and Innovation on Data Science* (S.-L. Peng, S.-Y. Hsieh, S. Gopalakrishnan, and B. Duraisamy, eds.), (Singapore), pp. 501–510, Springer Nature Singapore, 2021.

[13] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter,” *arXiv preprint arXiv:1910.01108*, 2019.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[19] H. Batra, N. S. Punn, S. K. Sonbhadra, and S. Agarwal, "Bert-based sentiment analysis: A software engineering perspective," in *Database and Expert Systems Applications* (C. Strauss, G. Kotsis, A. M. Tjoa, and I. Khalil, eds.), (Cham), pp. 138–148, Springer International Publishing, 2021.

[20] M. Hoang, O. A. Bihorac, and J. Rouces, "Aspect-based sentiment analysis using BERT," in *Proceedings of the 22nd Nordic Conference on Computational Linguistics* (M. Hartmann and B. Plank, eds.), (Turku, Finland), pp. 187–196, Linköping University Electronic Press, Sept.–Oct. 2019.