

Market Demand Analysis Using NLP in Bangla Language

by

Md Sabbir Hossain

18141007

Nishat Nayla

21341040

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
September 2021

© 2021. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:



Nishat Nayla
21341040



Md Sabbir Hossain
18141007

Approval

The thesis/project titled “Market Demand Analysis Using NLP in Bangla Language ” submitted by

1. Nishat Nayla (21341040)
2. Md Sabbir Hossain (18141007)

Of Summer, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on August 23, 2015.

Examining Committee:

Supervisor:
(Member)



Annajiat Alim Rasel
Lecturer
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Product market demand analysis plays a significant role for originating business strategies due to its noticeable impact on the competitive business field. Furthermore, there are roughly 228 million native Bengali speakers, the majority of whom use Banglish text to interact with one another on social media. Consumers are buying and evaluating items on social media with Banglish text as social media emerges as an online marketplace for entrepreneurs. People use social media to find preferred smartphone brands and models by sharing their positive and bad experiences with them. As a result, our goal is to gather Banglish text data and use sentiment analysis and named entity identification to assess Bangladeshi market demand for smartphones in order to determine the most popular smartphones by gender. We scraped data from social media with instant data scrapers and scraped data from Wikipedia with python web scrapers. Using Python's Pandas and Seaborn libraries, the raw data is filtered using NLP methods. To train our datasets for named entity recognition, we utilized Spacy's custom NER model, Amazon Comprehend Custom NER. A tensorflow sequential model was deployed with parameter tweaking for sentiment analysis. Meanwhile, we used the Google Cloud Translation API to estimate the gender of the reviewers using the BanglaLinga library. In this article, we use natural language processing (NLP) approaches and several machine learning models to identify the most in-demand items and services in the Bangladeshi market. Our model has an accuracy of 87.99 percent in Spacy Custom Named Entity recognition, 95.51 percent in Amazon Comprehend Custom NER, and 87.02 percent in the Sequential model for demand analysis. After Spacy's study, we were able to manage 80 % of mistakes related to misspelled words using a mix of Levenshtein distance and ratio algorithms.

Keywords: Market Demand Analysis, Sentiment analysis, Natural Language Processing, Name Entity Recognition, Tensor-flow, Gender Prediction, Banglish Text.

Acknowledgement

First of all we are very grateful to almighty Allah for whom we have successfully completed our thesis work without facing any serious difficulties. Then we show gratitude and respect towards our parents without their support and sacrifices we would not have been able to come this far. Further, we would like to show our heartfelt gratitude to our respected supervisor Annajiat Alim Rasel sir for his guidance, time, feedback and advice to make our thesis successful. We will be forever grateful to him for his kind support throughout the whole thesis process. Finally, a big thanks to Brac University for the necessary supplementary materials and resources to complete our work.

Table of Contents

Declaration	i
Approval	ii
Abstract	iii
Acknowledgment	iv
Table of Contents	v
List of Figures	vii
Nomenclature	viii
1 Introduction	1
1.1 Introduction	1
1.2 Problem Statement	2
1.3 Research Objectives	2
1.4 Obstacles	3
2 Background	4
2.1 Literature review	4
2.2 Previously Used Models	5
2.2.1 Naive Bayes Classification	5
2.2.2 TF-IDF algorithm	6
2.2.3 Word2Vec in Tensorflow	6
2.2.4 Fast Text Model	7
2.2.5 Support Vector Machines	8
2.2.6 Logistic Regression	9
2.2.7 K-Nearest Neighbor	10
3 Dataset Description	12
3.1 Data Collection	12
3.2 Data Preprocessing	12
3.2.1 Tweaking Phone List Algorithm	14
3.3 Data Train and Validation set	14

4	Research Methodology	15
4.1	Workflow of proposed model	15
4.1.1	Spell Correcter Algorithm For Entities	17
4.2	Approaches	17
4.2.1	Named Entity Recognition	17
4.2.2	Gender prediction	18
4.2.3	Sentiment Analysis	18
5	Algorithm Description	19
5.1	Amazon Comprehend Custom NER Model	19
5.2	Spacy NER Model	19
5.3	Sequential model of Tensorflow	20
6	Implementation	21
6.1	Spacy Custom NER Implementation	21
6.2	Amazon Comprehend NER Implementation	21
6.3	Sentiment Analysis	22
7	Result and accuracy	23
7.1	Spacy Custom NER Accuracy	23
7.2	Amazon Comprehend Accuracy	24
7.3	Sentiment Analysis Accuracy	24
7.4	Demand Analysis	24
8	Conclusion and future work	27
8.1	Conclusion	27
8.2	Future work	27
	Bibliography	30

List of Figures

2.1	Naive Bayes formula	6
2.2	Word2vec implementation	7
2.3	Fast text mode Implementation	8
2.4	SVM Graphing Method	9
2.5	Logistic Regression Formula with Graph	9
2.6	K-Nearest Neighbour visualization	10
3.1	Data Pre-processing Implementation	13
3.2	Phone list dataset	13
4.1	Flowchart of the Proposed Model	16
5.1	Amazon Comprehend Model	19
5.2	Spacy- NER Pipeline	20
5.3	Tensor Flow Sequential Model Layers	20
6.1	Spacy Custom NER Model Training Procedures	21
6.2	Sentiment Analysis Model Training Procedures	22
7.1	Spacy Custom NER Accuracy	23
7.2	Amazon Comprehend Accuracy	24
7.3	Amazon Comprehend Model output	24
7.4	Sentiment Analysis Accuracy	25
7.5	Male Female ratio pie chart	25
7.6	Product Demand Analysis Based on Gender	25
7.7	Product Demand Analysis Based on Female	26
7.8	Product Demand Analysis Based on male	26

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

CSV Comma Separated Values

LSTM Long Short-Term Memory

NER Name Entity Recognition

NLP Natural Language Processing

TFIDF Term Frequency Inverse Document Frequency

Chapter 1

Introduction

1.1 Introduction

To thrive in a competitive market, the rapidly changing nature of today's economic climate necessitated a good company plan. For start-ups, a persuasive business plan establishes perception and direction. If entrepreneurs wish to launch a new firm in this highly competitive industry, they must first obtain a thorough understanding of the current most gender-specifically demanded product. It is critical to be familiar with the interests of the target customers in order to engage them with a sophisticated approach. Besides, sentiment analysis is a sufficient method to observe consumers preferences, desired models and brands. Through sentiment analysis major data and information can be collected about consumers sentiment towards a brand or services. Sentiment analysis effectively obtains, quantifies, reclaim and analyzes consumers perceptions which benefits entrepreneurs to originate efficient business strategies. Nowadays, social media has become one of the biggest online marketplaces for potential buyers and sellers. It allows entrepreneurs to engage and correlate with the interests of the consumers and study their behaviours. Social media sentiments more intensively affect the stock market demand and price compared to traditional media [1]. Not only the stock market but also other business sectors are significantly influenced by social platforms. More than that, we emphasized on text since most of the Bengali people use Banglish text in social media for sharing their views and opinions. That is why we decided to analyze the Banglish text data from social media buy sell groups using natural language processing for a statistical and realistic market demand analysis for entrepreneurs based on smartphones. At first, we collected raw Banglish text data from social platforms buy and sell groups using automatic data scraping. After that, using natural language processing features, we filtered and clustered the data to turn into supervised data. Following, we applied machine learning algorithms (AWS Comprehend and Spacy NER) to train our dataset and used the Sequential model from tensorflow for performing sentiment analysis. Subsequently, validated accuracy of AWS Comprehend, Spacy NER and Sequential Model. These trained models and test datasets provide a proper idea about today's most demanding product based on gender and sentiment analysis. To sum up, we proposed a predictive model collecting raw data from social platforms, applying natural language processing and multiple data science algorithms to predict the market demand of smartphones based on consumer review for the entrepreneurs to have a crystal clear knowledge about the competitive market.

1.2 Problem Statement

Market demand analysis is a branch of business research that helps to find out the demanding product or service in a market and boost their operation in such a way to generate enough profits. People spend a lot of time on social media platforms. There are many buy and sell groups where people post any desirable item to buy. Often we can see some items are up for sale and many responses from customers. There are many items that cost way higher than people's budget. People seek those items within their affordable price range. Sometimes people bargain about items in the comments section. We can gather these data by web scraping and make a list of the items that appear on those groups. After listing those items, we can mark the importance of those goods by the meaning of comments and seeking posts. Finally, we can sort out the most demanding products of the market at that time which need more supply. Working with text is challenging, and here we are proposing to do it in the Bangla language. It is a language full of complex grammatical structures for words, sentences, and passages. A single word can be used in many affix items and can have different meanings based on its context. Also, we have to filter out relevant comments using natural language processing techniques. There is not any relevant work that has been previously done in product and service market demand analysis using NLP through Bengali language. For Bengali text understanding we will use various approaches including machine learning models, name entity recognition, sentiment analysis and BanglaLinga library. We will try to implement all of these models and find out the best solution to provide an accurate demand analysis model.

1.3 Research Objectives

This research aims to bring together the existing machine learning models and Natural language processing features to classify the current most demandable device in the market. The inputs will be users' posts and comments relevant to the device models, where the output will be the shortened list of the products that have highest demand in the current market. There has been vast work done in the field of stock price prediction but less amount of work has been done in the other sector's demand analysis. Also, our main focus is working on the Banglish text. Nowadays most of the people communicate in social media through Banglish text and for implementing an accurate model we choose Banglish texts to work with. Not only that but also our goal is to identify devices based on gender so that entrepreneurs can have a crystal clear idea to stock which product in which quantity. We will also include sentiment analysis to identify the positive reviewed devices. An entrepreneur can be successful if he knows the detailed market demand along with the positive and negative sides of the products. Finally, we will deploy the complete architecture as a private web application that will sort out the most demanding positively reviewed goods for the entrepreneurs.

1.4 Obstacles

Our main obstacle was Banglish text as it does not belong to any language. So we did not find any relevant dataset and it was challenging for us to make a supervised dataset based on Banglish text. There is also very little work that has been done in Banglish text so far. The comments we collected from social media were also full of irrelevant information and corrupted. So, sorting relevant accurate data was another challenge for us. Finally, gender prediction using name entity recognition was another difficult part as there are a lot of names that exist in Bangla language which belong to both genders male and female.

Chapter 2

Background

2.1 Literature review

Market demand analysis plays a significant role in the economic environment as it helps to determine business policies. Due to the gradual expansion of the business sector, market demand analysis has become a very popular research area. Hence, lots of previous research work has been done in this sector. In addition, using natural language processing to summarize sentiment analysis from a text has opened a new revolution in the field of machine learning. People used different approaches to natural language processing to analyze market demand and supply chain. To predict more accurately, scientific prediction methods are more beneficial [2]. Stock markets play a vital role in the economy but predicting and analyzing it is not an easy task as a lot of factors are involved with it. It is nearly impossible for the investors to keep track of all these factors, stocks as thousands of companies listed in a stock market. Although, a machine learning algorithm can easily record and analyze the data's keeping with all the significant factors. Therefore, text mining algorithms and Natural language processing can take input from authentic and inauthentic resources and make an analytical model for the stock market [3]. Early research on stock market prediction was mainly based on machine learning, support vector machines and so on. Now, the scenario has changed, investors are familiar with behavioural finance of the consumers. Sentiment analysis makes these procedures more efficient. A system developed by natural language processing techniques can facilitate the selection of the most dependable product in the market. Also, natural language processing can be used for classifying positive and negative data to provide a performance graph [27].

The effect of social platforms as a new emerging media on financial markets plays a significant role. It is proven that sentiment revealed through social platforms has a larger and longer-lasting impact on stock returns than views expressed in the news sources [1]. The sentiment of the public in various social platforms plays a significant role as a vital input in predictive model building for a stock price movement of an economical environment which can give a nearly close prediction. Comparing other machine learning and deep learning models with the social media public sentiment has the best ability to accurately analyze stock price movement [20]. The concept of sentiment analysis discusses the emotions, attitudes, and opinions of peoples. Applying multiple natural language processing features the exact percentage of positive, negative and neutral opinions can be driven. Thus, public opinions and

views on trending topics can easily be pursued by using sentiment analysis[19]. Entrepreneurs can be benefited in the process of making their business strategy by extracting and analyzing data from social networking sites. Large and microscopic companies are appearing in social networking sites to share their products and taking reviews from the customers. These companies use sentiment analysis to embrace the consumer interests and demands. This paper derives the accuracy of 85.25 % performing sentiment analysis using natural language processing in Twitter data [16].

Bangla language has significant impact on the communication process and digitization of this is very prominating. With a wide range of vocabulary, Bangla is a very complex language for embedding and clustering it's words. There is wealthy literature developing on word embedding strategies and there is plenty of scope of improvement in Bangla language processing [10]. Online engagement of Bangla language in the business sector is increasing day by day. For a product, there might be thousands of comments and reviews, it's hard for entrepreneurs to analyze all the comments and reviews to perceive the most demanding product. As people are more comfortable using their mother tongue for giving reviews, it is very rational to use sentiment analysis of positive and negative feedback written in Bangla language performing natural language processing and machine learning tools [24]. Nowadays the usage of acronyms and abbreviations on social media is increasing and this is broadly used by teenagers and young adults. Thus, the use of banglish text in social media is also growing. Banglish text is very flexible and a convenient way of communication for all kinds of people and online sellers and buyers are also communicating with this[29].

From the above discussion, it is clearly observed that most of the research in this field designed to predict stock price prediction and most of the Bangla language works are based on sentiment analysis. However, there are more prominent business strategies to analyze other than the stock market prediction. For instance, market demand-supply analysis, maintaining equilibrium curves, consumer sentiment analysis and so on. Furthermore, minor works based on consumer product review with sentiment analysis have been done in Banglish text. Banglish text sentiment analysis using natural language processing should be more prominent, and more research sectors should be included.

2.2 Previously Used Models

2.2.1 Naive Bayes Classification

To evaluate if user input has a positive or negative sentiment, a sentiment analysis system based on the Multinomial Nave Bayes classification method was constructed. Because it is simple and effective, the Nave Bayes algorithm is commonly used for categorization. When compared to other classic machine learning methods, the Multinomial Nave Bayes algorithm performs the best[21]. With some False positive and False negative data, the application of naive bayes can be presented[5]. Based on the Bayes theorem Naive Bayes classifier has been formulated as a probabilistic classifier. From the training data set this classifier constructs a classification model. Naive Bayes has a supervised learning mechanism which helps to classify any data into positive and negative. It describes the chances of occurring an event based on

some related conditions. It has independent pair of features[19]. Here, $P(c|x)$ gives us

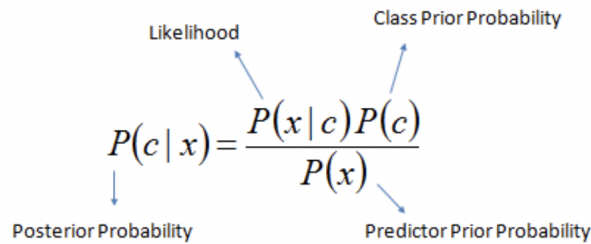
$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$


Figure 2.1: Naive Bayes formula

the probability of a class given predictor (x , attributes). $P(c)$ is the already known probability of class occurrence or selection. $P(x|c)$ points to the probability of the new predictor given class. $P(x)$ is the priorly affirmative

2.2.2 TF-IDF algorithm

TF-IDF algorithm is used to process data into numerical form. This algorithm is used to quantify words into documents. For this algorithm a computed weight has been assigned to each word, and that weight signifies the importance of the word in the document. This algorithm is previously used for text mining and retrieving information from text data. In term weighting, efficient feature extraction approaches such as the term frequency-inverse document frequency (TF-IDF) techniques are commonly used [17]. In terms of Bangla text document categorization, a model using a multilayer dense neural network with TF-IDF as the feature selection technique has a higher accuracy (84.58 percent) and F1 score (0.84) than the other well-known classification techniques [13]. A classifier with TF-IDF boosted features outperforms a combination of standard TF-IDF and state-of-the-art machine learning algorithms like LightGBM, Gradient Boosting, SVM, and others that use standard TF-IDF scores [12].

2.2.3 Word2Vec in Tensorflow

Tensorflow has official sentence embedding tools. This tool is vastly used in the NLP sector for clustering a random word. Millions of individuals post reviews and comments on social media every day, expressing their perspectives on a variety of topics. A method was suggested that combines sentiment analysis and classification. The major goal of this project is to provide an effective approach for providing a summary of Twitter users' interests based on their social textual data in five categories: sports, travel, fashion, food, and religion. The technique was tested and validated on a variety of publicly available corpora of Twitter reviews and comments. For user interest classification, the proposed method attained the highest accuracy of 97.3 percent [6].

Random Word	Words on cluster
আমরা	আমাদের, আমি, চাই, যখন, তাই, তারা, কি, সেই, কিন্তু, সবাই
তাঁর	তার, সেই, সাথে, তাঁদের, একজন, একই, ওই, তিনি, পরে, বলে
জন্য	প্রয়োজন, সুযোগ, জন্যে, পাশাপাশি, তাই, দরকার, কারণ, কিছু, কিন্তু, তাদের
কোন	কোনো, এমন, অন্য, কারণ, তবে, সেটা, বা, নেই, তাই, এখনো
পারে	হবে, পারবে, পারি, হলে, পারেন, হতো, চাই, চায়, পারেনি, থাকে
হতে	যেতে, থাকতে, করতে, হলে, না, রাখতে, তাই, তাহলে, তবে, দিতে
বড়	সবচেয়ে, অবস্থা, খুবই, খুব, আমাদের, অনেক, কিছু, মতো, মানুষের, আছে
টাকা	হাজার, লাখ, কোটি, টাকার, খরচ, পাঁচ, মাত্র, বিক্রি, তিন, প্রায়
নতুন	মাধ্যমে, তৈরি, কাজ, জন্য, বিভিন্ন, নানা, একটি, এই, সব, একই
দেখা	অদেখা, দেখাইত, দেখাক, দেখায়ই, দেখাসহ, দেখায়া, দেখাএইসবই, দেখাত, দেখাবো, দেখাবো

Figure 2.2: Word2vec implementation
[10]

2.2.4 Fast Text Model

The FastText package for word embedding and text categorization was built by Facebook’s AI Research team. FastText may be used to train both Skip gram and CBOW models. To test the prospect of employing Facebook’s open source fastText technology in apps on conventional computers and mobile systems, a new technique to addressing the classification problem has been introduced, notably because to its low processing power demand and faster model training. FastText’s method produced convincing results, exceeding typical machine learning-based classifiers in terms of accuracy [8]. A deep learning algorithm was devised, which successfully sorted unfavorable information on internet discussion sites into appropriate categories. Using the fastText word embedding technique, the suggested model used a text-based Convolution Neural Network (CNN) with word embedding. When compared to Word2Vec and the GLOVE model, fastText produced more efficient and accurate results. Toxic, Severe Toxic, Obscene, Threat, Insult, and Identity-hate are the six types of remarks classified by the model [18]. Instagram is a popular social media platform that could be used to spread hatred. For the Indonesian language, a study was conducted to detect hate speech in Instagram comments. As a classifier and word representation, they used FastText. The results of the experiment revealed that FastText outperformed Random Forest Decision Tree and Logistic Regression. FastText paired with the bigram function yielded the best results, with an F-measure of 65.7 percent [9].

Random Word	Words on cluster
আমরা	আমরা, কীআমরা, নয়আমরা, আমরাই, আমরাতো, হয়আ- মরা, কিআমরা, হোকআমরা, আমরা, লেআমরা
তঁর	তঁর, তঁরই, তঁরও, তঁহা, তঁরা, তঁতীও, তঁরাই, তঁ, তঁকৈসহ, ওঁর
জন্য	জন্য, জন্যে, জন্যও, জন্যে, সৌজন্য, জন্যে, এজন্য, জন্যই, জন্যি, এরজন্য
কোন	কোনস, কোন, কোনো, কোনোও, কোনো, কোনডা, কোন, কোনোই, কোনও, কোনই
পারে	পারো, পারে, পারেএ, পারেতখন, পারেঃ, পারেআর, পারেএমন, পারেনই, পারেনা, ঐপারে
হতে	ঝতে, লখনউতে, ধতে, ইইউতে, নড়তে, নতে, পেতে, চড়তে, অইতে, ওতে
বড়	বড়বড়, বড়র, হড়বড়, বড়ও, বড়ইর, বড়ছোট, ছোটবড়, ছোটছোট, ছোট, নড়বড়
টাকা	টাকা, দশটাকা, টাকায়ও, টাকাসহ, হাজারগুণ, দুইটাকা, হাজারও, দুটাকা, হাজারদিঘী, টাকাকী
নতুন	নতুননতুন, নতুনতর, নতুনা, নতুনই, নত, নতুন, জৈতুন, নিত্যনতুন, নতুনরা, চালু
দেখা	অদেখা, দেখাইত, দেখাক, দেখায়ই, দেখাসহ, দেখায়া, দেখাএইসবই, দেখাত, দেখাবো, দেখাবো

Figure 2.3: Fast text mode Implementation
[10]

2.2.5 Support Vector Machines

This is a machine learning algorithm which is a supervised learning model and associates with learning algorithms. Basic functionalities of a support vector machine is to analyze data to classify and perform regression analysis. A system is implemented in the study that offers a user with an authenticated score of any on-line product computed using the NLP approach known as Sentiment Analysis and Opinion Mining. They created a fine-grained Sentimental Analysis report using a Support Vector Machine (SVM) classifier. This classifier organizes the reviews received from the product's website URL into specified categories, allowing the user's thought process to be validated and a rating from 1 to 5 to be generated (1 being the lowest, 5 being the highest). This procedure allows any potential consumer to get a definitive judgment about the product evaluation [25]. The ability of TextCNN to extract text eigenvalues well and the classification function of Support Vector Machine (SVM) to find the maximum segmentation hyperplane were combined in a TextCNN-SVM combination model for text sentiment polarity classification that was implemented in the TensorFlow framework. The combined model's accuracy has increased by roughly 12.37 percent as compared to the traditional TextCNN model [30]. The use of an Enhanced SVM CTM-based technique is proposed for the analysis of data from Twitter tweets. Using NLP to perform the pre-processing and then using the SVM CTM technique to the filtered data for spam analysis. When compared to previous methodologies, the prediction ratio is superior [32]. Support vector machine effectively used in high dimensional spaces with a clear margin of

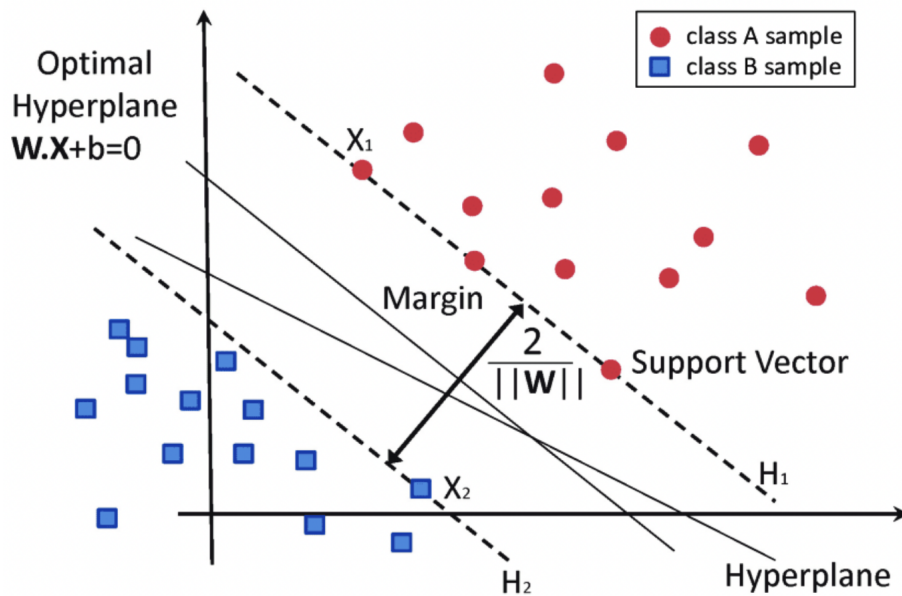


Figure 2.4: SVM Graphing Method [4]

separation. We can make the model efficient by tuning it's parameters.

2.2.6 Logistic Regression

In statistics, the logistic model is used to describe the probability of a certain class or event, such as pass/fail, win/lose, alive/dead, or healthy/sick, occurring. This may be used to mimic a number of scenarios, such as determining if a photograph contains a cat, dog, lion, or other animal. In today's world, the amount of data is continually increasing. Because everyone expresses themselves, shares their opinions, and even complains on social media, it is one of the factors contributing to the increase in data. The performance of A algorithm was evaluated using logistic regression. According to the results of the test, the average accuracy was 74 percent, with a 90:10 ratio of training and testing data. The accuracy level was unaffected by changing the folding quantity [7]. To improve sentiment classification, a mix of prior domain knowledge

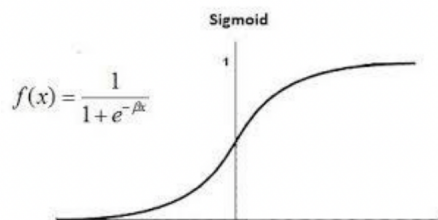


Figure 2.5: Logistic Regression Formula with Graph [22]

and logistic regression was used, as well as a gradient descent strategy to optimize the

modified logistic regression model. The approach is unique in that it incorporates previous domain information into the logistic regression model directly. Extensive experiments on a multi-domain sentiment data set were used to empirically evaluate the proposed approach. It was also compared to three baseline approaches and outperformed them all by a wide margin [26]. Covid-19 is a virus that has infected the entire world’s population and has had a significant impact on all parts of life. This may have a negative impact on mental health, thus a sentiment analysis study was done to learn more about mental health through public opinion on Twitter. The Logistic Regression approach will be used to classify Covid-19 tweets. The accuracy of the covid-19 tweets sentiment classification was 94.71 percent based on that research [23].

2.2.7 K-Nearest Neighbor

The k-nearest neighbors strategy, developed by Evelyn Fix and Joseph Hodges in 1951 and later extended by Thomas Cover in statistics, is a non-parametric classification approach. Data classification and regression are both done using it. In all cases, the input is the data set’s k closest training instances.

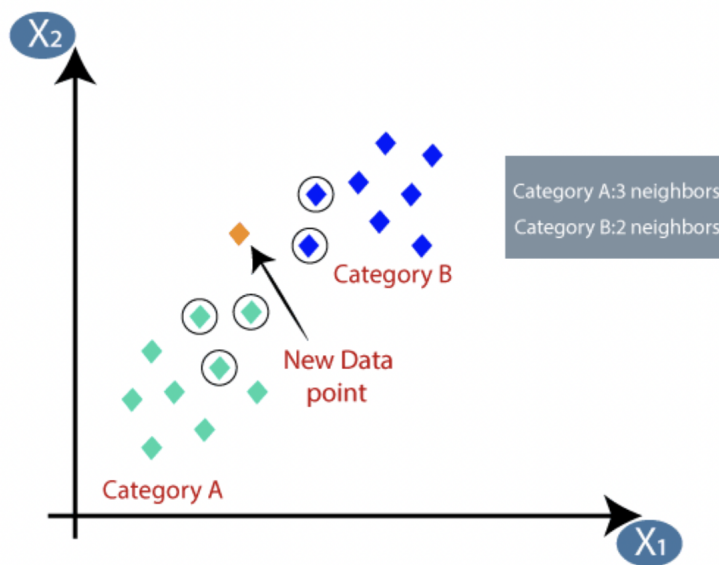


Figure 2.6: K-Nearest Neighbour visualization [31]

In a study, the K-Nearest Neighbor algorithm was used to categorize test data with $k = 3, 5, 7, 9$ and the Levenshtein Distance was utilized to correct inaccurate types. The goal of using a combination of the K-N Nearest Neighbor algorithm and Levenshtein Distance is to determine the level of accuracy. The accuracy of the K-Nearest Neighbor method is 50 percent for the values $k = 3$ and $k = 7$, whereas the accuracy of the K-Nearest Neighbor plus Levenshtein Distance algorithm is 65.625 percent for the value $k = 3$ [11]. A study looked at how statistical factors like interpacket arrival time, time to live, and quantity of packets may be used to analyze internet traffic and help minimize packet information invasion. Machine learning-based supervised classification algorithms such as Naive Bayes and K Nearest Neighbors were used

to automate the process of categorizing internet data. Experiments were conducted in order to achieve the maximum accuracy in identifying internet traffic based on transaction protocol. The results showed that the K-Nearest Neighbors algorithm has an accuracy of 85 percent, whereas the Nave Bayes algorithm has a maximum accuracy of 54 percent [14].

Chapter 3

Dataset Description

3.1 Data Collection

We made two datasets for training our model. One is based on the customer product review and the other is device list data from wikipedia. Banglish Texts do not belong to any particular language, it is a decomposed form of Bangla language with English alphabets. Similarly, there is not that much work has been done with Banglish Text reviews. For this reason, we could not find organized and supervised data related to customer product review. Nowadays Social networking sites play a phenomenal role in the producer - consumer relationship. There are a lot of mobile phone related groups or pages on social networking sites where consumers share their preferences, reviews. To get a proper analysis of product demand we decided to collect our raw data from social networking sites. Using the Instant Data Scraper tool we collected people's Banglish comments related to mobile phones from various social networking sites. We stored the unsupervised data in a CSV file for further analysis. And for our second dataset we collected smartphone model data from wikipedia using python web scraper. We collected raw data and made a supervised dataset using natural language processing features with manual interventions. Our dataset contains more than 10000 raw data collected from social media.

3.2 Data Preprocessing

We used an immediate data scraper to capture all of the information and saved it as several csv files. Multiple csv files were initially combined into a single large file. Then, based on column names, we examined the csv file and removed any superfluous columns. Only the name and comment columns were retained. Another issue was duplicate data, which was eliminated using Python's Pandas Library. We also got rid of any comments that didn't include a device model. The Seaborn library was then used to check for null values, which were then discarded using Python's Pandas module. For sentiment analysis, we manually labeled the datasets and utilized regular expressions with edit distance methods to identify entities.

Any device's name must be known in order for it to recognize device names from comments. As a result, we used pandas to compile a list of cellphones from Wikipedia. Then, using a loop, we went through the data and chose the columns we required. We also included Apple devices in the mix. Finally, we created a CSV file with our phone list. The phone list datasets have to be fine-tuned. The phone model includes

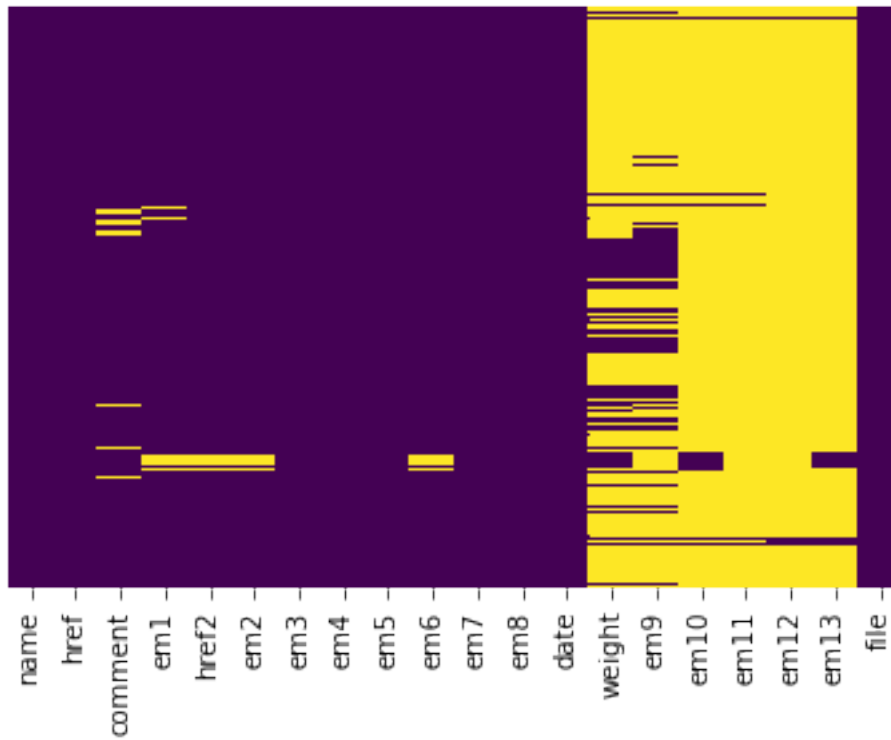


Figure 3.1: Data Pre-processing Implementation

4	Asus PadFone mini	Asus	2013/12	Android 4.1 "Jelly Bean"	[5]
5	Asus PadFone E	Asus	2014/01	Android 4.1 "Jelly Bean"	[6]
6	Asus PadFone Infinity Lite	Asus	2014/02	Android 4.1 "Jelly Bean"	[7]
7	Asus ZenFone 5 (2014)	Asus	2014/04	Android 4.1 "Jelly Bean"	[8]
8	Asus ZenFone 4 (2014)	Asus	2014/05	Android 4.1 "Jelly Bean"	[9]
9	Asus ZenFone 6 (2014)	Asus	2014/05	Android 4.1 "Jelly Bean"	[10]
10	Asus PadFone S/X	Asus	2014/06	Android 4.4 "KitKat"	[11][12]
11	Asus ZenFone 4.5 (2014)	Asus	2014/07	Android 4.4 "KitKat"	[13]
12	Asus ZenFone 5 (2014)	Asus	2014/07	Android 4.4 "KitKat"	[14]
13	Asus PadFone X mini	Asus	2014/10	Android 4.4 "KitKat"	[15]
14	Asus Pegasus (X002)	Asus	2014/12	Android 4.4 "KitKat"	[16]
15	Asus ZenFone 5 (2015)	Asus	2015/01	Android 4.1 "Jelly Bean"	[17]
16	Asus ZenFone 2	Asus	2015/03	Android 5.0 "Lollipop"	[18][19][20]
17	Asus ZenFone 3	Asus	2016/08	Android 6.0 "Marshmallow"	[21][22]
18	Asus ZenFone 4	Asus	2017/10	Android 7.0 "Nougat"	[23]
19	Asus ZenFone Max Plus M1	Asus	2017/11	Android 7.0 "Nougat"	[24]
20	Asus ZenFone Max M1	Asus	2018/05	Android 8.0 "Oreo"	[25][26]
21	Asus ZenFone Max Pro M1	Asus	2018/05	Android 8.0 "Oreo"	[27]

Figure 3.2: Phone list dataset

the brand name as well as the model number, as can be seen. The device model "Galaxy S20" was, for example, "Samsung Galaxy S20" in the dataset. People, on the other hand, do not leave comments or tweets containing the manufacturer's name. To solve this problem, we used a sophisticated approach depicted in the diagram below to remove the developer's name from the device model.

3.2.1 Tweaking Phone List Algorithm

```
for  $i$  in phonelist do
  phonemodel += regex delete [0-9]* (str(phonelist.model[i])) with none
  phonemodel += regex delete /. * (str(phonelist.model[i])) with none
  phonemodel += regex delete .* (str(phonelist.model[i])) with none
  if regex search (developer name, model, ignoring case) then
    if length (remove developer name from model name) > 6 then
      remove developer name from model name
    end if
  end if
end for
```

We utilized regular expressions to eliminate certain indicators and the manufacturer's year from the data, as well as the developer's name. The device's model, on the other hand, had to be at least 7 characters long, including whitespaces. If the developer name's character length falls below 7, we didn't cut it off. As a consequence, instead of "Apple iPhone XS" , we got "iPhone XS."

3.3 Data Train and Validation set

We splitted our data into a train set, test set and validation set. Our train set contains 3300 pieces of data. For refining the train set we dropped the rows which don't contain any device name. We normalized the dataset to its lower case. Further we used the stemming technique to noun entity using the edit distance algorithm and levenshtein ratio because there are not any predefined stemmers for Banglish corpora. Data labeling is performed using regular expression matches considering the case. Annotated sets were required for the Spacy Custom NER and Amazon Comprehend Custom NER models. We made two functions to make an annotated train set from a regular train set so that we can feed this to Spacy NER and Amazon Comprehend custom named entity detection classifier. After that, we labeled each comment manually. Finally our train and validation set is ready to fit and check the accuracy of our model.

Chapter 4

Research Methodology

4.1 Workflow of proposed model

This section represents our proposed work which focuses on a strategy to sentiment analysis on public facebook groups, page's comments data and their gender prediction for detecting most demanding device entities. The architectural overview describing an overall process design of sentiment analysis, gender prediction and named entity recognition which is shown in Fig. 4.1. The developed method is based on several parts that are data scraping from social media, scrapping valid product name entities from authentic sites, pre-processing of the extracted social-media data using Natural Language toolkits and Regular Expressions. To begin our research we need the clean device names list. To get that, firstly after scraping data from authentic sources we preprocessed those using pandas, manual coding and regexmatch and generated a phone list csv file so that we may use this in our research. To pre-processing and simplify the datasets firstly we have dropped all unnecessary columns and kept only the name and comments column. Secondly, we visualized the dataset with Seaborn heatmap and dropped all null and duplicate values. We need a trainset to train sentiment analysis classifiers and named entity recognition classifiers. We splitted the datasets into train, test and validation sets ensuring each set contains at least 3000 valid data. Further, we labeled the sentiment manually for the training set so that we may get better results. After that, we labeled the name entity from comments. We defined a function to match the device names by comparing between the phone list and the comments data while ignoring the case and slight spelling errors with a combination of levenshtein ratio and edit distance algorithms. For predicting the gender, we have implemented a pre-trained model from BanglaLinga. As the predefined model doesn't support Banglish names, Banglish text has been translated into Bangla names. Most of the pretrained models support gender detection from Bangla text. Although, there are not enough labeled Banglish datasets to predict gender from a Banglish name. Thus we found a feasible solution for our problem to translate Banglish to Bangla using the Google Cloud Translation api.

We implemented the endpoint in our notebook successfully. Later, we feed the Banglish text into the pre-trained gender prediction model and it successfully gives the desired output. Though it has some exceptions like, it can only predict accurately with one word. So we had to send the prefix of the name. Consequently, we got too many exceptions like names starting with 'Md', 'Phd', 'Dr', 'Mrs', 'Miss', 'Engr'

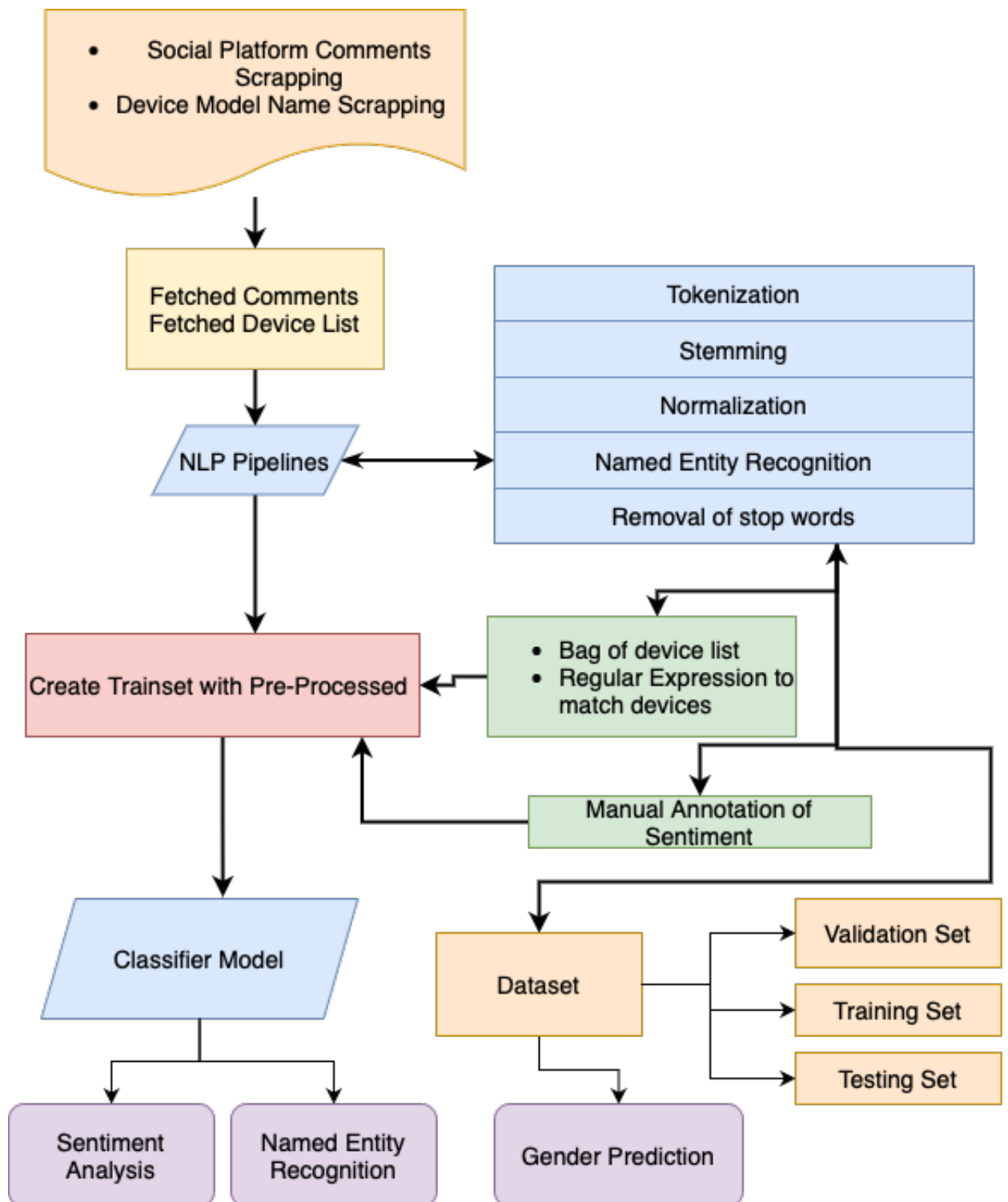


Figure 4.1: Flowchart of the Proposed Model

etc. We had to handle those exceptions using some cases with the help of regular expressions. Lastly, we have the clean train set and validation set. We made our sentiment analysis classifier with the Sequential model from keras. Then, we trained the model with the manual annotated train set of 5000 data with dropout value of 0.25. After fitting the model we moved to train our named entity classifier. Firstly we trained the NER model from Spacy. We fine tuned the parameters and got a satisfactory result. Secondly we used the Amazon Comprehend for custom named entity recognition. We trained the Comprehend with our labeled annotated sets and got more satisfactory results. Finally there are some misspelled device keywords which both models predict as incorrect. We made a combination of Levenshtein ratio and Edit Distance algorithms to correct the misspelled predicted devices name. Finally we plotted the most demanding device list based on the gender in the current market.

4.1.1 Spell Correcter Algorithm For Entities

```

for i in df.iterrows(): do
    comment = str(df.comment[index])
    DeviceModels = wordTokenize(comment)
    for x in range(len(list3)): do
        MinDistance = 100
        MaximumMatch = DeviceModels[x]
        HighestRatio = 0
        for y in range(len(ModelName)): do
            d=editdistance.eval(DeviceModels[x], str(ModelName[y]))
            e=LevenshteinRatio(DeviceModels[x], str(ModelName[y]))
            if d < 3 && d >= 0 && d < MinDistance && e > 0.55 && e >
                HighestRatio then
                MinDistance = d
                HighestRatio= e
                MaximumMatch=str(ModelName[y])
            end if
        end for
        DeviceModels[x] = MaximumMatch
    end for
    df.coment[index] = " ".join(DeviceModels)
end for

```

4.2 Approaches

4.2.1 Named Entity Recognition

The objective of named-entity recognition is to find and categorize named entities stated in unstructured text into predefined categories such as human names, organizations, places, medical codes, time expressions, quantities, monetary values, percentages, and so on. We used the Spacy custom NER and Amazon Comprehend custom named entity recognition models to recognize the device entities from the extracted comments.

4.2.2 Gender prediction

Product demand differs from one individual to the next. Preferences are generally divided into two categories based on gender. It's critical to understand product demand depending on gender. Bangla Linga's gender prediction model was implemented.

There were other flaws, such as the model's inability to correctly identify gender based on entire names. We simply utilized the first name to solve this problem. Despite the fact that first names such as 'Md,' 'Phd,' 'Dr,' 'Mrs,' 'Miss,' 'Engr,' and so on have been used in a variety of circumstances, We used regular expression to handle these exceptions and passed the appropriate first name to the gender prediction model. The algorithm also fails to detect gender from Banglish text names. To resolve the issue, we had to get the Google Cloud Platform and implement the cloud translation api. This cloud translation endpoint performed better than most other translation api in python. Finally we come up with the gender prediction part properly.

4.2.3 Sentiment Analysis

Sentiment analysis is a part of natural language processing which is also known as data mining. It extracts subjective information from a text and categorizes it into positive or negative. A sequential model from tensorflow has been used for sentiment analysis. This model is best fitted for a plain stack of layers where each layer has exactly one tensor input and output. We set the pad sequences max length to 300, spatial dropout 1D to 0.25 and a dropout value for LSTM is 0.5. In our training we used the adam optimiser. For training the neural network model we chose the sigmoid activation function. It gives us about 89% accuracy.

Chapter 5

Algorithm Description

5.1 Amazon Comprehend Custom NER Model

Natural language processing (NLP) is used by Amazon Comprehend to extract information from documents. Any text file in UTF-8 format is processed by Amazon Comprehend. By recognizing entities, important words, language, attitudes, and other common components in a document, it produces insights. Amazon Comprehend returns a list of entities identified in a document, such as persons, places, and locations. With the argument of device name location in string, we created the annotated train set. To train the classifier, we had to send the train, test and

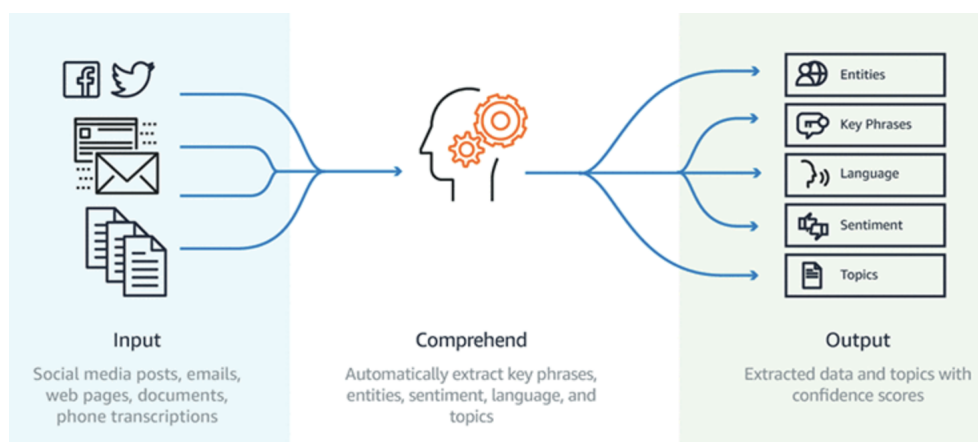


Figure 5.1: Amazon Comprehend Model [15]

validation set using json format. The output also needs to be decoded from json. Our trained Amazon Comprehend model gives us an accuracy of 91 % in detecting device names.

5.2 Spacy NER Model

A component for recognizing named entities based on transitions. Non-overlapping labelled spans of tokens are identified by the entity recognizer. The transition-based approach encodes some assumptions that are successful for "conventional" named entity recognition tasks, but may not be appropriate for other span identification

problems. The loss function, in particular, optimizes for entire entity correctness,

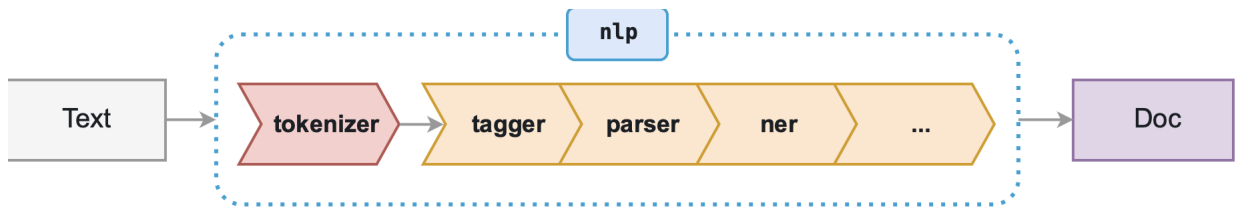


Figure 5.2: Spacy- NER Pipeline

thus if inter-annotator agreement on boundary tokens is low, the component will likely do badly on our issue. The transition-based method also expects that the most important data about your entities will be around their initial tokens.

5.3 Sequential model of Tensorflow

In Keras, the simplest approach to create a model is sequential. It allows you to layer-by-layer construct a model. Each layer contains weights that match the weights of the one above it.

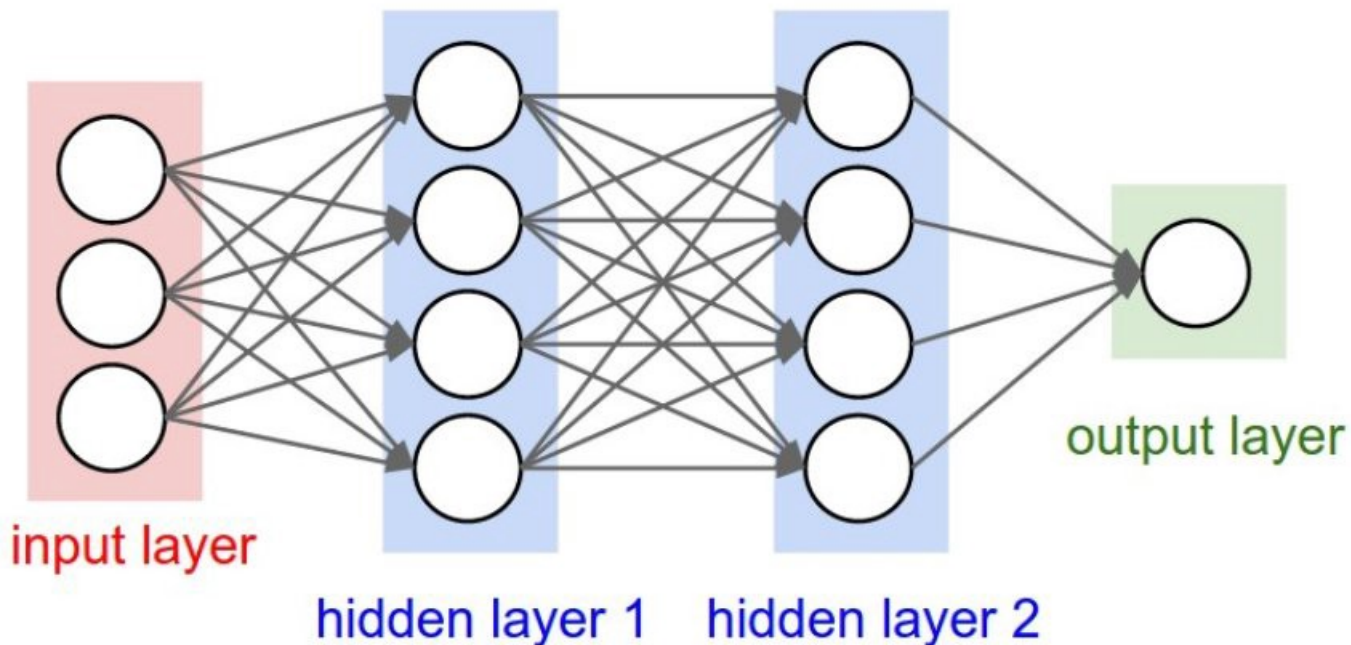


Figure 5.3: Tensor Flow Sequential Model Layers
[28]

Chapter 6

Implementation

6.1 Spacy Custom NER Implementation

Spacy has the 'NER' pipeline component that identifies token spans fitting a pre-determined set of named entities. For fitting our NER model we choose Spacy's custom NER model with minibatch compound. For training purposes, it needs annotated labeled data. We made a function to make a trainable dataset for Spacy NER from a regular train set. We feed around 3300 data with its industrial level annotation. We chose the mini batch iteration of Spacy custom NER value of 5 and drop value of 0.1 and trained our custom model. It took several minutes to train in Google Colab.

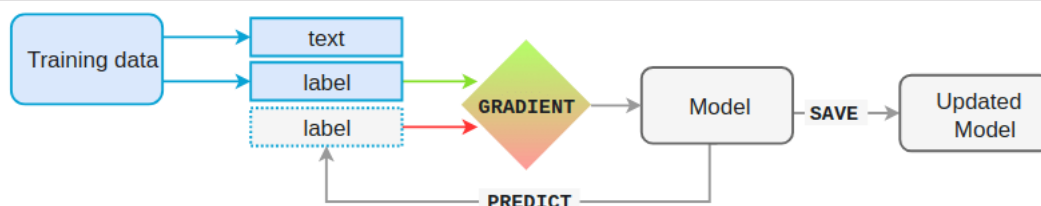


Figure 6.1: Spacy Custom NER Model Training Procedures [33]

6.2 Amazon Comprehend NER Implementation

We have also used AWS's built in custom named entity recognition model. We had train the model with lines of tweets and another CSV file containing the annotation for that lines of comments to train. We made a function to convert the trainset to both a text file which contains lines of comments and the CSV file which is the annotated documentation for the text file. We passed around 3380 labeled datasets and trained the custom model within the web interface. We chose around 10% of traindata during the training phase to check the precision of the model.

6.3 Sentiment Analysis

To analyze the demand from the comments or tweets we need to pass the text from a machine learning technique. We implemented the Sequential model from Keras in Tensorflow. We used the Adam Optimizer and around 4 epochs with the pad sequence of 300 to train our model with around manual labeled 4650 comments.

Train on 3720 samples, validate on 930 samples

Epoch 1/4

3720/3720 [=====] - 87s 23ms/sample - loss: 0.1009 - val_accuracy: 0.9817

Epoch 2/4

3720/3720 [=====] - 96s 26ms/sample - loss: 7.5561e-94 - val_accuracy: 0.9817

Epoch 3/4

3720/3720 [=====] - 85s 23ms/sample - loss: 3.6895e-79 - val_accuracy: 0.9817

Epoch 4/4

3720/3720 [=====] - 86s 23ms/sample - loss: 2.4045e-41 - val_accuracy: 0.9817

Figure 6.2: Sentiment Analysis Model Training Procedures

Chapter 7

Result and accuracy

7.1 Spacy Custom NER Accuracy

We had a validation test set which consists of around 2000 labeled data. We matched every specific value with the Spacy NER result and got an outstanding accuracy of 87.99 percent. However we implemented some methods to fix spelling errors to increase the accuracy.

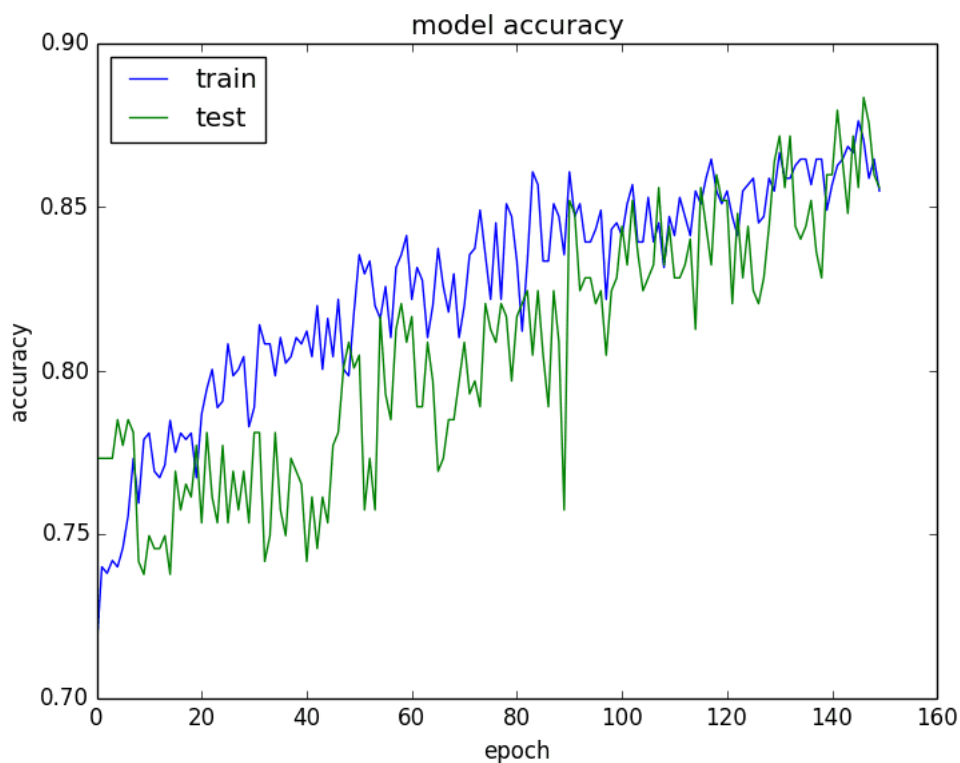


Figure 7.1: Spacy Custom NER Accuracy

7.2 Amazon Comprehend Accuracy

Amazon Comprehend Custom Named Entity Recognition depicts an outstanding F1 score of 95.66. To test the model we implemented the test set in similar fashion

Entity types performance			
Custom entity type	F1 score	Precision	Recall
DEVICE	95.66 %	95.51 %	95.81 %

Figure 7.2: Amazon Comprehend Accuracy

only without the annotation. It took around a few seconds to test 2000 data using our custom trained model and output the result as JSON format shown in Fig 7.3. We can see that it performs better than the Spacy custom NER system with an 92.3% accuracy but it is not cost effective.

```
{"Entities": [{"BeginOffset": 0, "EndOffset": 10, "Score": 0.9999977946351262, "Text": "realme 25s", "Type": "DEVICE"}], "File": "test.txt", "Line": 142}
{"Entities": [{"BeginOffset": 22, "EndOffset": 31, "Score": 0.9999985694906286, "Text": "oneplus 8", "Type": "DEVICE"}], "File": "test.txt", "Line": 143}
{"Entities": [{"BeginOffset": 0, "EndOffset": 6, "Score": 0.9999983310727032, "Text": "iphone", "Type": "DEVICE"}], "File": "test.txt", "Line": 144}
{"Entities": [{"BeginOffset": 4, "EndOffset": 12, "Score": 0.999995231630692, "Text": "realme 8", "Type": "DEVICE"}], "File": "test.txt", "Line": 145}
{"Entities": [{"BeginOffset": 5, "EndOffset": 14, "Score": 0.9999954700730597, "Text": "oneplus 8", "Type": "DEVICE"}], "File": "test.txt", "Line": 146}
```

Figure 7.3: Amazon Comprehend Model output

7.3 Sentiment Analysis Accuracy

We tested our sentiment analysis model and found the peak floating number at where it differs from the positive or negative demand. We tuned the parameter for many times and tested with around 2000 datas and found the accuracy of 86.02 % shown in the Fig 7.4

7.4 Demand Analysis

We need to see the results after a significant workload. We created a pie chart that shows the proportion of men and women that use social media and their desire for gadgets. We can see that about 53% of men and 47% of women are interested in this sector. We created a bar chart based on the gender ratio of the top 8 most popular phones in the current Bangladeshi market, and we can see that the Realme 8 is the most popular phone among both genders.

In the bar chart below, we've forecasts the top three preferred phones among males and females.

Likewise, by monitoring public tweets, we can forecast many sorts of items and goods demands.

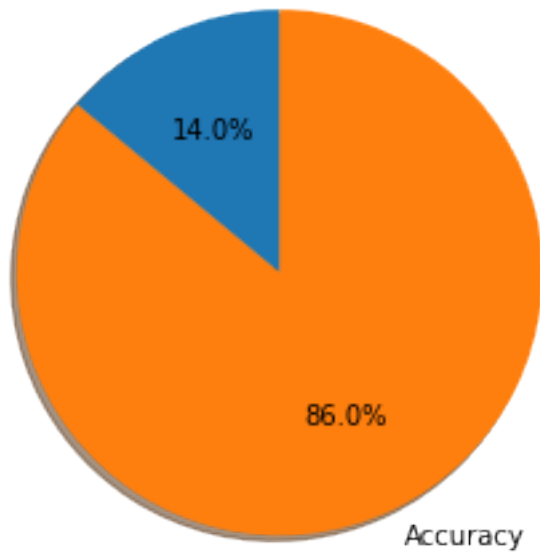


Figure 7.4: Sentiment Analysis Accuracy



Figure 7.5: Male Female ratio pie chart

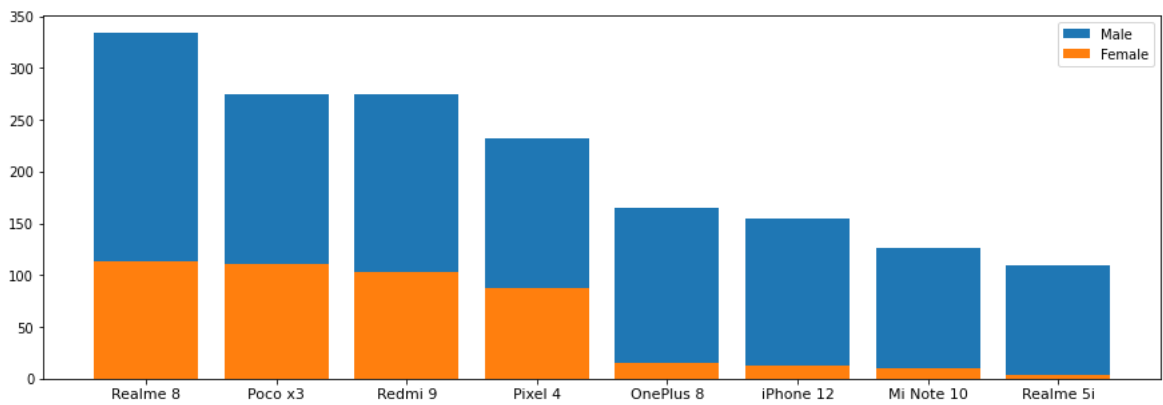


Figure 7.6: Product Demand Analysis Based on Gender

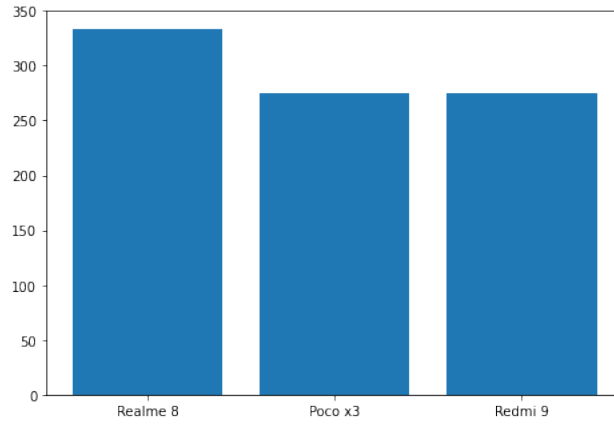


Figure 7.7: Product Demand Analysis Based on Female

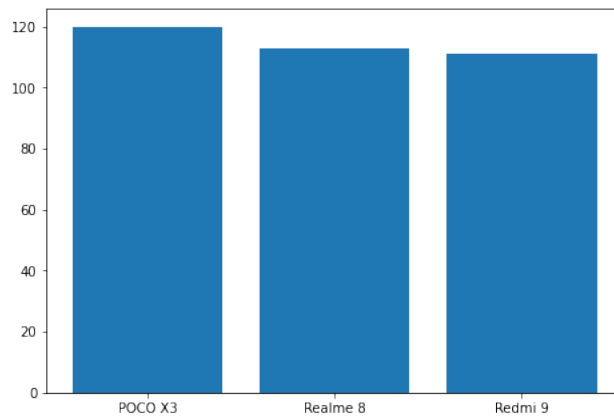


Figure 7.8: Product Demand Analysis Based on male

Chapter 8

Conclusion and future work

8.1 Conclusion

In this paper, we used several machine learning algorithms such as the Spacy NER Model, Amazon Comprehend Custom Entity Recognition model. Then we used a Sequential model from keras for name entity recognition and we predicted gender from Bangla Linga library. Then we train and test our model with our customized data set. Through the name entity recognition model we successfully identified the gender of the person based on their names. Our model successfully identified the most demandable device model names from consumers' comments and posts data. Amazon's comprehensive model gives 95.66 % accuracy on our data set and Spacy custom NER model gives 92.3 % accuracy. Then we performed sentiment analysis and it 86.02 % accurately predicted the positive reviewed devices. Finally, our project provides a graphical representation of most demanding and positively reviewed devices from current market scenarios based on gender.

8.2 Future work

Currently we are working on only social platforms data but further we will also include data from e-commerce websites, newspapers and other sources. Also we are collecting data manually but further if we get proper sponsors then we will add auto data collecting features to collect more data. Then we will also focus on other factors such as desired price, device color, condition, camera preferences, gaming performance into the demand analysis field so that our model can provide a nearly accurate demand prediction.

Bibliography

- [1] H. Chen, P. De, Y. Hu, and B.-H. Hwang, “Sentiment revealed in social media and its effect on the stock market,” pp. 25–28, 2011. DOI: 10.1109/SSP.2011.5967675.
- [2] H. Zongxiang, “Research on the prediction of product demand under market economy,” vol. 1, pp. 153–156, 2011. DOI: 10.1109/ICIII.2011.41.
- [3] S. Abdullah, M. Rahaman, and M. Rahman, “Analysis of stock market using text mining and natural language processing,” pp. 1–6, May 2013. DOI: 10.1109/ICIEV.2013.6572673.
- [4] E. Garcia-Gonzalo, Z. Fernández-Muñiz, P. J. Garcia Nieto, A. Sánchez, and M. Menéndez, “Hard-rock stability analysis for span design in entry-type excavations with learning classifiers,” *Materials*, vol. 9, p. 531, Jun. 2016. DOI: 10.3390/ma9070531.
- [5] M. Mertiya and A. Singh, “Combining naive bayes and adjective analysis for sentiment detection on twitter,” vol. 2, pp. 1–6, 2016. DOI: 10.1109/INVENTIVE.2016.7824847.
- [6] A. H. Ombabi, O. Lazzez, W. Ouarda, and A. M. Alimi, “Deep learning framework based on word2vec and cnnfor users interests classification,” pp. 1–7, 2017. DOI: 10.1109/SCCSIT.2017.8293054.
- [7] W. Ramadhan, S. Astri Novianty, and S. Casi Setianingsih, “Sentiment analysis using multinomial logistic regression,” pp. 46–49, 2017. DOI: 10.1109/ICCEREC.2017.8226700.
- [8] H. A. Chowdhury, M. A. H. Imon, and M. S. Islam, “Authorship attribution in bengali literature using fasttext’s hierarchical classifier,” pp. 102–106, 2018. DOI: 10.1109/CEEICT.2018.8628109.
- [9] N. I. Pratiwi, I. Budi, and I. Alfina, “Hate speech detection on indonesian instagram comments using fasttext approach,” pp. 447–450, 2018. DOI: 10.1109/ICACSSIS.2018.8618182.
- [10] Z. Sultana Ritu, N. Nowshin, M. Mahadi Hasan Nahid, and S. Ismail, “Performance analysis of different word embedding models on bangla language,” pp. 1–5, 2018. DOI: 10.1109/ICBSLP.2018.8554681.
- [11] N. Anggraini and M. J. Tursina, “Sentiment analysis of school zoning system on youtube social media using the k-nearest neighbor with levenshtein distance algorithm,” vol. 7, pp. 1–4, 2019. DOI: 10.1109/CITSM47753.2019.8965407.
- [12] U. Bhattacharjee, P. Srijith, and M. S. Desarkar, “Term specific tf-idf boosting for detection of rumours in social networks,” pp. 726–731, 2019. DOI: 10.1109/COMSNETS.2019.8711427.

- [13] M. Chakraborty and M. Nurul Huda, “Bangla document categorisation using multilayer dense neural network with tf-idf,” pp. 1–4, 2019. DOI: 10.1109/ICASERT.2019.8934530.
- [14] M. Dixit, R. Sharma, S. Shaikh, and K. Muley, “Internet traffic detection using naïve bayes and k-nearest neighbors (knn) algorithm,” pp. 1153–1157, 2019. DOI: 10.1109/ICCS45141.2019.9065655.
- [15] given-i=Amazon, “Analyze content with amazon comprehend and amazon sagemaker notebooks,” Apr. 23, 2019. [Online]. Available: <https://aws.amazon.com/blogs/machine-learning/analyze-content-with-amazon-comprehend-and-amazon-sagemaker-notebooks/>.
- [16] M. R. Hasan, M. Maliha, and M. Arifuzzaman, “Sentiment analysis with nlp on twitter data,” pp. 1–4, 2019. DOI: 10.1109/IC4ME247184.2019.9036670.
- [17] A. I. Kadhim, “Term weighting for feature extraction on twitter: A comparison between bm25 and tf-idf,” pp. 124–128, 2019. DOI: 10.1109/ICOASE.2019.8723825.
- [18] S. Mestry, H. Singh, R. Chauhan, V. Bisht, and K. Tiwari, “Automation in social networking comments with the help of robust fasttext and cnn,” pp. 1–4, 2019. DOI: 10.1109/ICICT1.2019.8741503.
- [19] V. Sathya, A. Venkataramanan, A. Tiwari, and D. D. P.S., “Ascertaining public opinion through sentiment analysis,” pp. 1139–1143, 2019. DOI: 10.1109/ICCMC.2019.8819738.
- [20] J. Sen and S. Mehtab, “A robust predictive model for stock price prediction using deep learning and natural language processing,” *Proceedings of the 7th International Conference on Business Analytics and Intelligence (BAICONF, 2019), Indian Institute of Management, Bangalore, INDIA,*, Dec. 2019. DOI: 10.36227/techrxiv.15023361.v1.
- [21] G. P. Wiratama and A. Rusli, “Sentiment analysis of application user feedback in bahasa indonesia using multinomial naive bayes,” pp. 223–227, 2019. DOI: 10.1109/CONMEDIA46929.2019.8981850.
- [22] GeeksforGeeks, “Advantages and Disadvantages of Logistic Regression,” Sep. 2020. [Online]. Available: <https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/>.
- [23] Imamah and F. H. Rachman, “Twitter sentiment analysis of covid-19 using term weighting tf-idf and logistic regresion,” pp. 238–242, 2020. DOI: 10.1109/ITIS50118.2020.9320958.
- [24] M. A. Shafin, M. M. Hasan, M. R. Alam, M. A. Mithu, A. U. Nur, and M. O. Faruk, “Product review sentiment analysis by using nlp and machine learning in bangla language,” pp. 1–5, 2020. DOI: 10.1109/ICCIT51783.2020.9392733.
- [25] A. Shetty, D. Makati, M. Shah, and S. Nadkarni, “Online product grading using sentimental analysis with svm,” pp. 1079–1084, 2020. DOI: 10.1109/ICICCS48265.2020.9121098.
- [26] N. A. Wasi and M. Abulaish, “Document-level sentiment analysis through incorporating prior domain knowledge into logistic regression,” pp. 969–974, 2020. DOI: 10.1109/WIAT50758.2020.00148.

- [27] R. D. E. Al., “Stock prediction by using nlp and deep learning approach,” *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 1S, pp. 202–211, 2021. DOI: 10.17762/turcomat.v12i1s.1611.
- [28] E. . Allibhai, “Building a deep learning model using keras - towards data science,” Jun. 2, 2021. [Online]. Available: <https://towardsdatascience.com/building-a-deep-learning-model-using-keras-1548ca149d37>.
- [29] S. Das, M. S. Islam, and I. Mahmud, “A deep learning study on understanding banglish and abbreviated words used in social media,” pp. 1690–1695, 2021. DOI: 10.1109/ICICCS51141.2021.9432339.
- [30] X. Dong, R. Hu, Y. Li, M. Liu, and Y. Xiao, “Text sentiment polarity classification based on textcnn-svm combination model,” pp. 325–328, 2021. DOI: 10.1109/ICAICA52286.2021.9498156.
- [31] javatpoint, “K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint,” 2021. [Online]. Available: <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning>.
- [32] V. Shrivastava, S. Karsoliya, B. Verma, and N. K. Gupta, “Social data analysis: Cyber recruitment analysis spam detection over twitter dataset using svm amp; arima model,” pp. 1–7, 2021. DOI: 10.1109/ICAECT49130.2021.9392543.
- [33] Spacy, “Training Pipelines Models · spaCy Usage Documentation,” 2021. [Online]. Available: <https://spacy.io/usage/training/>.