NORTH SOUTH UNIVERSITY

##### DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING



### **CSE499R: DIRECTED RESEARCH**

#### SENTIMENT ANALYSIS USING PRODUCT REVIEWS

#### AND CUSTOMER INTERACTIONS

**Project Report**

**Group Members**

**Name:** SADMAN KABIR **Name:** MARUF HASAN ARNOB

**ID:** 1811987042 **ID:** 1812701642

**Email:** sadman.kabir@northsouth.edu **Email:** maruf.arnob@northsouth.edu

**Section:** 15

**Team Number:** 19

**Submission Date:** 19th June 2023

Under of Supervision of

**Dr. Dihan Md. Nuruddin Hasan**

Assistant Professor

Department of Electrical & Computer Engineering

North South University

Dhaka, Bangladesh

Spring 2023

**Contents**

|  |  |
| --- | --- |
| **Declaration** | **i** |
| **Approval** | **ii** |
| **Abstract** | **iv** |
| **Acknowledgements** | **x** |
| **Glossary** | **xi** |
| **1 Introduction** | **1** |
| 1.1 Demand for Insights into Public thought and Behavior . . . . . . . . | 1 |
| 1.2 The Beginning . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 2 |
| 1.3 Our Sentiment Analysis Overview . . . . . . . . . . . . . . . . . . . . . | 3 |
| **2 Sentiment Analysis and Its Use In ML Applications** | **6** |
| 2.1 Analysis of Twitter emotions by means of NLP methods . . . . . | 6 |
| 2.2 Sentiment analysis on Twitter with MNB and Logistic Regression | 6 |
| 2.3 Classification of Twitter Sentiment Using Deep Neural Networks . . .. . | 7 |
| 2.4 A Sentence-Level Sentiment Analysis of Online News and Blog Content  Using Machine Learning . . . . . . . . . . . . . . . . . . . . . . . . . . . | 7 |
| 2.5 Customer reviews posted online inside the hospitality industry’s dining  establishments . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 8 |
| 2.6 Using machine learning, we do a sentiment analysis of IPL tweets. . .. . | 9 |
| 2.7 Using Machine Learning and Public Opinion to Predict the Price of Bitcoin  in Real Time . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 9 |
| 2.8 Negation and Speculation Detection in Sentiment Analysis Using Machine  Learning . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 10 |
| 2.9 Using machine learning for sentiment analysis in business intelligence. . | 10 |
| **3 Problem Statement** | **11** |
| 3.1 Limitaions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. . | 11 |
| **4 Use Cases of Sentimental Analysis** | **12** |
| 4.1 Monitoring of the brand . . . . . . . . . . . . . . . . . . . . . . . . . . . | 12 |
| 4.2 Competitive Research . . . . . . . . . . . . . . . . . . . . . . . . . . .. . | 12 |

v

vi *CONTENTS*

* 1. [Monitoring for fires and giving priority to service to customers 13](#_TOC_250041)
  2. [Evaluation of the product 13](#_TOC_250040)
  3. [Research on the market and observations of tendencies in the industry 13](#_TOC_250039)
  4. [Analytics of the workforce and monitoring of employee engagement 14](#_TOC_250038)

1. [Literature Review 15](#_TOC_250037)
   1. [Categorization 15](#_TOC_250036)
   2. [Bag of Opinions Model 15](#_TOC_250035)
   3. [Neural Networks 16](#_TOC_250034)
   4. [Studies 16](#_TOC_250033)
2. [Methodology and Research Design 17](#_TOC_250032)
   1. [Collection of Datasets and its Features 17](#_TOC_250031)
   2. [Extraction of sentiment phrases and POS tagging 17](#_TOC_250030)
   3. [Identification of Negative Expressions 19](#_TOC_250029)
   4. [The real-world labels 19](#_TOC_250028)
   5. [Formation of Feature Vectors 20](#_TOC_250027)
3. [Technologies That Are In Use 22](#_TOC_250026)
   1. [Python 22](#_TOC_250025)
   2. [Stanford CoreNLP 22](#_TOC_250024)
   3. [Beautiful Soup 22](#_TOC_250023)
   4. [Scikit-learn 23](#_TOC_250022)
4. [Sentiment Classification Algorithms 24](#_TOC_250021)
   1. [Linear Regression 25](#_TOC_250020)
   2. [Logistic Regression 26](#_TOC_250019)
   3. [Decision Tree 26](#_TOC_250018)
   4. [Support Vector Machine (SVM) 27](#_TOC_250017)
   5. [Naive Bayes 29](#_TOC_250016)
   6. [KNN 30](#_TOC_250015)
   7. [K-means Clustering 30](#_TOC_250014)
   8. [Random Forest 30](#_TOC_250013)
   9. [Dimensionality Reduction Algorithms 31](#_TOC_250012)
   10. [Gradient Boosts and Ada Boost 32](#_TOC_250011)
5. [Results and Discussions 33](#_TOC_250010)
   1. [Classification of Multiclass 33](#_TOC_250009)
   2. [Binary classification 34](#_TOC_250008)
   3. [Performing an effectiveness analysis on the BERT model 35](#_TOC_250007)

CONTENTS vii

* 1. [Confusion Matrix and Accuracy 36](#_TOC_250006)
  2. [A plot showing the receiver’s operating characteristics 37](#_TOC_250005)
  3. [Methods of Evaluation 38](#_TOC_250004)
  4. [Result on manually-labeled sentences 39](#_TOC_250003)
  5. [Result on machine-labeled sentences 39](#_TOC_250002)
  6. [Review-level categorization 40](#_TOC_250001)

1. [Contribution of Individual Team Member 42](#_TOC_250000)

**List of Figures**

* 1. Rating system of Amazon. . . . . . . . . . . . . . . . . . . . . . . . . . . 4
  2. Dataset Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4
  3. Sentiment Polarity Categorization Process. . . . . . . . . . . . . . . . . . 5
  4. System Diagram 21
  5. Bar Chart showing Frequency of Ratings 24
  6. Bar Chart for Accuracy, Precision, Recall, and Roc-Auc 25
  7. Logical Regression Classifier Model 26
  8. Decision Tree Classifier Model 27
  9. Decision Tree Algorithm 27
  10. Linear SVM Model 28
  11. SVM Mapping 29
  12. Naive Bayes Classifier Equation 29
  13. Random Forest Tree 31

viii

**List of Tables**

* 1. Top 10 sentiment phrases based on occurrence 18
  2. Statistical information for word tokens 19
  3. Table of Accuracy, Precision, Recall and Roc Auc 36
  4. Confusion Matrix for Multinomial NB 41
  5. Confusion Matrix for Logistic Regression 41
  6. Confusion Matrix for Liner SVC 41
  7. Contribution of Individual Team Member 42

ix

1. **Introduction**
   1. **Demand for Insights into Public thought and Behavior**

In making choices, "what people think" has always been an essential piece of information for the vast majority of us. There was a time when the majority of people did not have access to the Internet; at that time, many of us relied on word of mouth to find a reliable auto mechanic, learn about the candidates running for local office, learn about potential employees through reference letters, and choose between dishwashers. However, thanks to the Internet and the World Wide Web, we can now learn about the perspectives and experiences of the large number of individuals who are neither our personal acquaintances nor well-known professional reviewers. Likewise, an increasing number of people are sharing their thoughts online with complete strangers. So say two polls of over 2,000 persons in the United States.

We’d like to stress that people’s reasons for looking for information or sharing their thoughts online go beyond the purchase of products and services. The requirement for knowledge of political matters is also crucial. For instance, Rainie and Horrigan surveyed over 2500 American individuals and analyzed the 31 percent of Americans, or over 60 million people, who used the internet to learn about the 2006 elections and communicate their opinions with others throughout that year.

One explanation for this increase in interest in novel systems that deal directly with views as a first-class object is the aforementioned data’s revelation of users’ insa- tiable appetite for and dependence on online advice and recommendations. Horrigan found that while the vast majority of American internet users had a favorable experience conducting product research online, 58% of those same people found the available online information to be inadequate. Therefore, it is imperative that superior information- access technologies be developed to benefit product and information consumers. Vendors of goods and services are beginning to pay greater attention to the opinions expressed by consumers online because of the interest shown by individuals and the potential effect of these thoughts. What follows is a snippet from a whitepaper that provides an example of the potentials that are being discussed.

Now more than ever, thanks to the proliferation of Web 2.0 tools like blogs, message boards, P2P networks, and other kinds of social media, customers have a platform from which to voice their thoughts and feelings about any company, good or bad. Market leaders are learning that the opinions of their customers on social media can have a

1

1. *CHAPTER 1. INTRODUCTION*

significant impact on the loyalty, purchasing decisions, and advocacy of their customers and the customers of their competitors.

Whether it’s for public relations, to catch instances of fraud3, or to gather intel on the competition, marketers have long had a vested interest in keeping tabs on media coverage of their companies. However, traditional monitoring approaches have been ren- dered ineffective as a result of the media fragmentation and shifting consumer behavior. According to Technorati, there are around 75,000 new blogs established every day, with a total of 1.2 million new postings. Clipping services, field agents, and impromptu stud- ies are examples of [conventional] strategies that can’t keep up with the current rate of change.

Because of this, businesses aren’t the only ones that may benefit from systems that can automatically analyze consumer mood, as expressed largely in online forums.

* 1. **The Beginning**

There has been a consistent undercurrent of interest in the field of sentiment analysis and opinion mining for quite some time, even before the recent great explosion of academic effort. Projects based on early views might be considered pioneers in the field. Metaphors, narrative, point of view, emotion, evidentiality in text, and related subjects dominated later research. There have been hundreds, if not thousands, of articles written about the research challenges and possibilities presented by sentiment analysis and opinion mining since around 2001. It’s important to note the following causes for this "land rush:"

* + - The development of more sophisticated techniques for using machine learning in NLP and IR;
    - The proliferation of the Internet and, in particular, the rise of review aggregation websites have increased the accessibility of datasets for training machine learning algorithms.
    - Acknowledgment of the interesting intellectual difficulties and business and intelli- gence applications the region provides.

*1.3. OUR SENTIMENT ANALYSIS OVERVIEW* 3

* 1. **Our Sentiment Analysis Overview**

The feeling gives rise to an attitude, conviction, or conclusion, which we refer to as emotion. Analysis of people’s sentiments, also known as opinion mining, delves into the thoughts and feelings that people have on specific topics. The internet is a wonderful resource for acquiring information about emotive topics. Users have the ability to post their own content through a variety of social media platforms, such as online forums, microblogs, and social networking sites. A great number of social media networks make their application programming interfaces (APIs) publicly available, which encourages researchers and developers to gather data and conduct analyses. For instance, the three primary API versions that are currently provided by Twitter are the REST API, the Search API, and the Streaming API. Using the REST API, developers may acquire information on users and their statuses; using the Search API, they can look for specific content on Twitter; and using the Streaming API, they can collect content that is being posted to Twitter in real-time. In addition, developers may mix such APIs to create apps that are unique to their needs. As a consequence of this, sentiment analysis gives the impression of having a strong basis, which is made possible by the assistance of large volumes of online data.

Sentiment analysis might be more challenging with these internet data sources due to a number of factors. The first problem is that people are permitted to submit their own material, therefore the validity of people’s opinions cannot be guaranteed. For instance, online spammers post spam on forums rather than engaging in genuine debate. While some spam is absolutely useless, some spam contains misinformation or fake opinions. The availability of the ground truth for such internet data is the second problem. A ground truth is more akin to a label placed on an opinion, designating whether it is favorable, unfavorable, or neutral.

The information that was used in this investigation originated from a database consisting of Amazon product reviews that were obtained throughout the months of February and April of 2014. The following two strategies have helped to mitigate the effects of the aforementioned problems to some degree: Before being published, every review of a product has to go through moderation first. Second, it is necessary to give each review a score that may be taken into consideration for the overall outcome. The ranking was determined using a technique based on a star scale, with 5 stars being the highest rating and 1 representing the lowest rating.

4 *CHAPTER 1. INTRODUCTION*



Figure 1.1: Rating system of Amazon.

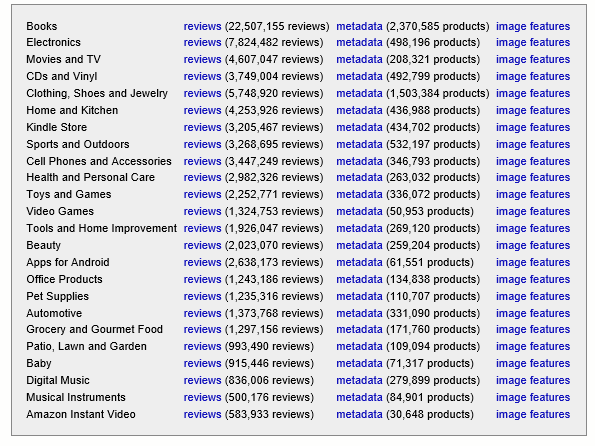


Figure 1.2: Dataset Summary

In this paper, one of the fundamental problems in sentiment analysis—the cat- egorization of sentiment polarity—is investigated. The structure of this paper’s overview

*1.3. OUR SENTIMENT ANALYSIS OVERVIEW* 5

as well as the proposed approach for categorization are both depicted in flowchart form in Figure 2, which may be found below. The majority of our contributions are made during Phases 2 and 3. phase two: 1. A strategy for recognizing negation phrases is proposed and put into practice; 2. A method for computing sentiment scores is proposed; and

3. An approach for generating feature vectors is presented for classifying the polarity of sentiments. the third step is: 1. Two different categorization experiments for sentiment polarity are carried out, one based on the phrase level and the other based on the review level; 2. The efficiency of three different classification models is evaluated and compared using the experimental data.

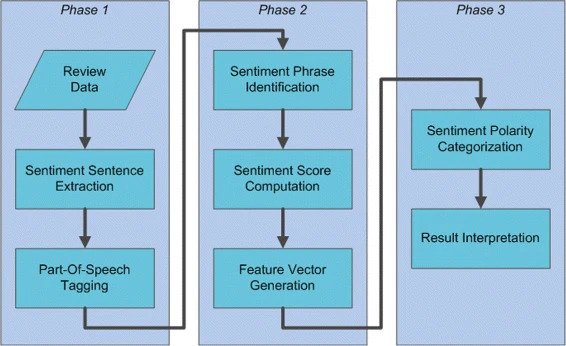


Figure 1.3: Sentiment Polarity Categorization Process.

The following portions of this document are structured as follows: In the section under "Background and literature review," we briefly discuss some significant sentiment analysis studies. Section "Methods" presents the computer program and classification models that were employed in this investigation. Our detailed sentiment analysis tech- niques are available in the section under "Background and literature study." The section "Results and remarks" contains the experiment’s findings. The section under "Review- level classification" contains discussion and planned work. The article is concluded in the "Conclusion" section.

1. **Sentiment Analysis and Its Use In ML Applications**
   1. **Analysis of Twitter emotions by means of NLP methods**

A massive amount of information gleaned from Twitter opinions is required for opinion mining. In natural language processing, numerous ways assist extract tweets from Twitter directly. There is no coherence to the tweets. It requires to analyze and sanitize tweets to achieve opinion mining, structured data. All links, hashtags, capitalized terms, repeated phrases, short-form terms, spelling errors, special symbols, Twitter characters, and resid- ual content are removed from the data before the research is conducted. Extraction and conversion of text to the data frame, removal of text URLs, removal of stop words (the, a, etc.), deletion of usernames and profiles, deletion of numbers and unneeded spaces, deletion of dots, and conversion of Emojis from Latin to ASCII are all examples of data cleaning. Removing data often means stripping tweets of their content. It merely have the text of tweets after processing and cleaning. When a user clicks on a word in a tweet, the Vader lexicon will pull up its definition from WordNet. An increasing word’s value is assessed and indicated as a tweet emotion score. If opinions are reached, it labels every tweet as favorable, normal, and bad by applying a machine learning classifier.

* 1. **Sentiment analysis on Twitter with MNB and Logistic Regression**

On Twitter, sentiment analysis can be contested. Some tweets are widely understood in rude language, and some really brief phrases show just the barest traces of emotion, which presents a problem. Hashtags, URLs, abbreviations, emojis, and acronyms are all often used on Twitter. A variety of machine-learning techniques are used to achieve high precision in areas such as tweet prediction, data extraction, and table layout. Using machine learning techniques, such as Multinomial Naive Bay and the logistic regression

6

*2.3. CLASSIFICATION OF TWITTER SENTIMENT USING DEEP NEURAL NETWORKS*7

algorithm using the train data on the test outcomes, the algorithms are put through their paces. The author takes into account the airline sentiment data set and the IMDB analysis data collection. Using the Count Vectorizer function in machine learning, people of all backgrounds may achieve successful outcomes. The author notes that the Logistic Regression with Count Vectorizer test set results to exhibit the following characteristics.

* 1. **Classification of Twitter Sentiment Using Deep Neural Networks**

The tweets are categorized using ML techniques as Multilayer Propon (MLP), Naive Bayes, Fuzzy Classification, Decision Tree, and Support Vector Machines (SVM.). These methods allow for a more nuanced evaluation of the assessment’s link to each individual piece by comparing and contrasting many component vectors with a designated class. The accuracy, time to notify, alert threshold, and F calculation are only some of the performance metrics examined in the Twitter dataset. There is an assessment of these methods. For categorization techniques, the accuracy may be anywhere from 72.66 to

92.34 percent, the precision anywhere from 72.16 to 90.12 percent, the recall anywhere from 72.81 percent to 94.34 percent, and the F-measurement anywhere from 72.4 to 92.3 percent. The results of the Twitter Sentiment analysis indicated that SVM was superior to all other methods.

* 1. **A Sentence-Level Sentiment Analysis of On- line News and Blog Content Using Machine Learning**

Two metrics used often in pre-processing and posting analytics are applicable to the detection and assessment of refusals at the phrase level in news items and forums. Both

8 *CHAPTER 2. SENTIMENT ANALYSIS AND ITS USE IN ML APPLICATIONS*

the BBC News Report and the multimodal dataset were utilized in the authors’ research. Data cleansing, removing numbers and making documents all lowercase, stroke, and frequency are only some of the methods used. Opinion and syuzhet packages, along with a dictionary strategy and lexical procedures, are used to identify the speaker’s emotional state. More criteria, such as positive term, negative word, down the tower, amplifier, de-amplifier, adverse combination, etc., affect the polarity of each line, making sentiment cluster superior at the sentence level. Here, we employ the Naive Bayes (NB) and Vector Support Machine (SVM) classification methods. The accuracy of Naive Bayes is 96.64 percent, while that of Vector Machine is 94.16 percent with the use of algorithms. When evaluating the probabilities and characteristics of each word in the provided phrases, Naive Bayes produces better results than the SVM.

* 1. **Customer reviews posted online inside the hospitality industry’s dining establishments**

To gather data, purify data, convert data, and reduce data using the string-to-Word Vector in order to assess the efficacy of many supervised machine learning methods and to discover if opinion mining can be extended to uncover phony positive and negative reviews. Classification algorithms are used to categorize data in a dataset. The study employs a number of different classifiers, including Naive Bayes, Decision Tree, Support Vector Machine, K-NN, and K-star. When both datasets are put through five different supervised machine learning algorithms or classifications, a confusion matrix is produced for each method. As a consequence of their analysis, the authors find that the SVM method outperforms both the k-nearest neighbor and the random forest algorithms in terms of precision, accuracy, negative detection rate, down rate, F1 metric, error rate, and runtime for both datasets.

*2.6. USING MACHINE LEARNING, WE DO A SENTIMENT ANALYSIS OF IPL TWEETS.*9

* 1. **Using machine learning, we do a sentiment analysis of IPL tweets.**

For example, offer a social media opinion mining method using this very machine learning strategy. With the use of an algorithm that takes into account what the hashtag IP TEAM is, users of Twitter’s API services may browse a list of tweets related to the 2016 Indian Premier League. Random Forest methods are compared to preexisting supervised machine learning algorithms in terms of their accuracy, precision, and sensitivity.

* 1. **Using Machine Learning and Public Opin- ion to Predict the Price of Bitcoin in Real Time**

Figure out where the value of a bitcoin will go against the dollar in the near future using machine learning and opinion mining. Extracted tweets were evaluated for their adherence to machine learning principles, and the relationship between tweets and Bitcoin price variations was followed by IT. It investigated several algorithms for masterminding with supervised learners in order to build a prediction model and produce a comprehensive analysis of future market rates. It’s also tough to create accurate forecasts due to how tricky it may be to pin down the presence of a reliable model of Time Series (ARIMA). Rather, LSTM Recurrent Neural Networks, which use long-term memory cells (RNN). Then, we compared the expected future cost of bitcoin to that predicted by the more traditional approach (ARIMA) using long-term memory (LSTM) approaches and found that the former was more accurate. Since the ARIMA RMSE model is 209.263, the multi- function LSTM is the most accurate method, with RMSEs of 198.448 (mono feature) and

197.515 (several feature), respectively.

10 *CHAPTER 2. SENTIMENT ANALYSIS AND ITS USE IN ML APPLICATIONS*

* 1. **Negation and Speculation Detection in Sen- timent Analysis Using Machine Learning**

The identification of negative knowledge and the detection of speculation play the most important role in sentence analysis. Machine learning will enable this sort of emotional analysis. Incorporating this method within the explanation of the text’s polarity improves the analysis. There are two phases to the analysis; in the first, problems and ambiguity are identified. Second, the whole calculation of this signal is performed. This feature, which is useful in opinion mining, is the brainchild of the Simon Fraser University Review corpus. Methods that are more rigorous in their approach, use more sophisticated lan- guage analysis tools for automating the evaluation of viewpoints, and put in more effort in the information extraction process are better able to spot potentially damaging data. Subjective domain labeling improves polarity detection. Given the subjective nature of conjecture. The final percentages for f1 inference during hint identification and f1 nega- tion were 92.37 and 89.64 respectively. During f1 scope detection tasks, they achieved a denial rate of 84.07 percent and an inference rate of 78.88 percent. And the PCRS for distrust was 71.43 percent, while it was 80.26 percent for rejection. According to the results of the study, detecting the scope of a term by automatic detection of its cues requires both lexical data and syntactic traits.

* 1. **Using machine learning for sentiment anal- ysis in business intelligence**

Categorization using sentiment analysis as a useful tool for assessing textual data gath- ered from various web sources. Emotion analysis is just another data mining method that puts machine learning to the test on textual information. Finding, assessing, and aggre- gating the wide variety of user opinions, reviews, comments, and ideas accessible online is crucial to better decision making. Opinion mining has the potential to significantly affect corporate choices since it provides a real-time, accurate, and dynamic picture of customers. Further study opportunities in Business Intelligence were also identified.

1. **Problem Statement**

An application that gathers and analyzes user evaluations of certain products. The reviews would be divided into favorable and unfavorable ones. The firms might utilize the unfavorable evaluations to improve their products depending on consumer input. The application also lists the benefits and drawbacks of each specific product feature. The program will also give reports on the results of the sentiment analysis done on the goods. Additionally, we want to develop a system for recommending items to consumers based on their feature needs.

* 1. **Limitaions**

Every system has certain flaws, or to put it another way, every system in the world oper- ates within some set parameters. Our system essentially operates within four constraints.

* + 1. Sarcasm: It is quite challenging for a machine to comprehend the precise meaning of a sarcastic evaluation of any product. Even humans occasionally have trouble deciphering some of these snarky remarks.
    2. Errors in Grammar: People frequently make spelling, punctuation, and gram- mar mistakes as a result of social networking apps. Most of the time, customers purpose- fully enter incorrect spellings to convey their views about the product. This makes it challenging for the computer to determine the precise meaning behind a customer re- view.
    3. Repetition of Words: In our method for reviewing products, we concentrate on nouns that appear in sentences and check to determine if they are features of the product using our feature database. However, customers frequently substitute pronouns for the correct feature names when expressing their opinions on a product. In these instances, the system is unable to identify the pronoun as a statement of fact since the feature name is not present in the product in question.

11

1. **Use Cases of Sentimental Anal- ysis**

The practice of opinion mining is utilized in a variety of fields. In spite of the fact that the areas of application for sentiment analysis are connected to one another, they all improve efficiency by monitoring shifts in public opinion.

* 1. **Monitoring of the brand**

Analysis of your company’s emotional connection to its products or services enables you to monitor your market credibility, identify existing or future reputational issues, and act swiftly.

* 1. **Competitive Research**

Capable of observing and analyzing how society evaluates competitors, just as you eval- uate how they contribute to any firm, you must have this ability. What kinds of things do other industry leaders anticipate receiving the most from their customers? Was there something that the other companies didn’t have or did wrong? Which mediums of com- munication do customers prefer when interacting with other businesses? Make use of this information to develop the contact, marketing campaigns, and services in general, as well as to give services and goods to clients.

12

* 1. *MONITORING FOR FIRES AND GIVING PRIORITY TO SERVICE TO CUSTOMERS*13

## Monitoring for fires and giving priority to service to customers

Using emotional grouping is something that businesses in the hospitality industry, bank- ing institutions, grocery stores, transportation, and other industries are doing in order to make the most of their work with customers. Text analysis users also have the ability to automate the process of polarity classification, style classification, dimension classifica- tion, and priority classification for incoming customer support communications. As the flame has to be stoked before it can become a fire, the fresh messages coming in from customers who are the least content and the most irate are being analyzed first.

## Evaluation of the product

Successful businesses create what’s known as a minimum viable product (MVP), get early input, and continue to improve a product even after it’s been released to the public. Surveys, social media and online forums, as well as personal encounters with customer service, all provide data. This data ocean presents a number of challenges, including the determination of which consumer groups to query, the assessment and categorization of comments, and so on.

## Research on the market and observations of tendencies in the industry

The problem of retrieving enormous volumes of unstructured data is solved via the analy- sis of emotions. Advertisers are able to watch and investigate patterns of client behavior in real-time utilizing these text analyses, which enables them to foresee potentially harmful actions and provide management with assistance in making educated decisions.

14 *CHAPTER 4. USE CASES OF SENTIMENTAL ANALYSIS*

## Analytics of the workforce and monitoring of employee engagement

The analysis of SA tech employee surveys is automated by specialists, which assists reso- lution specialists in resolving issues and complaints in a more timely manner. Personnel managers are able to identify and keep track of the overall number of comments, the outcomes of groups broken down by departments, and keywords, as well as track how the feelings of employees have evolved over time. Analysis of workers’ sentiments enables the next stage of mood management by providing capabilities for real-time monitoring.

# Literature Review

## Categorization

In order to categorize movie ratings as either "high" or "poor" based on its reviews, SVM and Nave Bayes classifiers are trained. From the textual assessments, several linguistic characteristics are retrieved, and feature selection is carried out using TF-IDF and infor- mation gain. The most accurate SVM classifier, according to the findings, used features that were chosen based on information gain. The model, however, lacks granularity since it cannot differentiate between "poor" (reviews with a rating of 2 stars) and "worst" (reviews with a rating of 1 star).

On the other side, a model that uses sentiment dictionaries to predict a review’s star rating was given. The study uses the unigram model to represent text, like the ma- jority of polarity determination methods. Due to the unigram model’s frequent inability to accurately capture phrase patterns, polarity incoherence results. The authors also use an n-gram model to get around this flaw. But because text vectors are represented in this manner, enormous, inefficient sparse matrices known as n-gram sparsity bottlenecks are produced.

## Bag of Opinions Model

The Bag of Opinions model is presented in order to address the issue of polarity inco- herence. The unigram and n-gram models’ shortcomings are addressed by this model. The root words, modifiers, and negation words make up its three parts. Using the ridge regression approach, each opinion from the corpus of reviews is given a score. The sum of all the independent scores from each opinion is then used to compute the final score.

15

16 *CHAPTER 5. LITERATURE REVIEW*

## Neural Networks

Neural networks have been employed for sentiment categorization in another investiga- tion. Along with considering written evaluations and user input, it also considers the characteristics of the product. A review, according to the writers, cannot be generalized to all users because different people may have different impressions and opinions about it. The results show that the constructed model performs better than other classification techniques like SVM and HAN, especially when there is a lack of sufficient data.

## Studies

Several studies have also been carried out to show how sentiment analysis may be used in the travel and hospitality industry. We discovered, for instance, that there is a link between customer reviews and hotel ratings. To ascertain the polarity of the evaluations, the authors analyzed the hotel reviews and tested the Naive Bayes classifier against a word lexicon. To forecast the ratings based on guest feedback and additional hotel characteristics like cost and location, a linear regression model was created.

# Methodology and Research De- sign

## Collection of Datasets and its Features

We utilized a 5-core Amazon review dataset. The selected dataset includes reviews of variety of products that were purchased from Amazon.com. It has 11 characteristics and 1,128,437 rows, which are described below. Each row represents a customer review and contains the following feature variables:

* reviewerID - ID of the reviewer
* asin - ID of the product
* reviewerName - name of the reviewer
* vote - helpful votes of the review
* style - a dictionary of the product metadata
* reviewText – customer review text
* overall - rating of the product
* summary - summary of the review
* unixReviewTime - time of the review (unix time)
* reviewTime - time of the review (raw)
* image - images that users post after they have received the product.

## Extraction of sentiment phrases and POS tagging

According to Pang and Lee’s recommendations, sentiment analysis should only be uti- lized with content that is subjective. In the course of our research, all of the subjective material was eliminated and stored away for subsequent examination, rather than the objective stuff being deleted. The subjective content consists entirely of statements with an emotional tone. A statement is considered to be sentimental if it contains at least one word that might be construed either positively or negatively. First, each statement was

17

18 *CHAPTER 6. METHODOLOGY AND RESEARCH DESIGN*

|  |  |  |
| --- | --- | --- |
| **Phrase** | **Type** | **Occurrence** |
| not worth | NOA | 26329 |
| not go wrong | NOA | 15446 |
| not bad | NOA | 15122 |
| not be happier | NOA | 14892 |
| not good | NOA | 12919 |
| don’t like | NOV | 42525 |
| didn’t work | NOV | 38287 |
| didn’t like | NOV | 21806 |
| don’t work | NOV | 10671 |
| don’t recommend | NOV | 9670 |

Table 6.1: Top 10 sentiment phrases based on occurrence

broken down into its component English words using a process called tokenization.

Every word in a phrase serves a certain syntactic function, which explains how that word should be used in the sentence. The syntactic roles are often referred to by another term, which is the components of speech. There are eight different parts of speech that make up the English language. These are the verb, noun, pronoun, adjective, adverb, preposition, conjunction, and interjection. In the field of natural language processing, part-of-speech taggers, often known as POS taggers, have been developed to classify words in accordance with their respective parts of speech. When it comes to doing sentiment analysis, a point-of-sale tagger proves to be particularly useful for two reasons, which are as follows: 1) The emotions that are connected to pronouns and nouns aren’t always present in real life. It is feasible to use a POS tagger to differentiate between words that may be used in a number of various parts of speech, which enables it to filter out sentences of this sort. This is made possible by the fact that it is possible to use several parts of speech to refer to the same word. For instance, the word "improved" may inspire a different amount of sensation when it is employed as a verb as compared to when it is used as an adjective. This is because of the distinct roles that each of these words play in the sentence. This experiment made use of a max-entropy version of a point-of-sale tagger that had been built for the Penn Treebank Project. The fact that the tagger is able to discriminate between 46 distinct tags suggests that it is capable of detecting syntactic functions that are more complicated than the mere 8 that it is currently limited to. As an illustration, all of the verb tags that have been added to the point-of-sale tagger are detailed in Table 1, which can be located further down this page.

A Python program that can run in parallel was created in order to increase the pace of tagging given the vast volume of texts. Because adjectives, adverbs, and verbs are the words that primarily express feeling, there are over 25 million adjectives, over 22 million adverbs, and over 56 million verbs labeled out of all the sentiment phrases.

* 1. *IDENTIFICATION OF NEGATIVE EXPRESSIONS* 19

Token Type Mean Median Positive Word Token 3.18 3.16

Negative Word Token 2.75 2.71

Table 6.2: Statistical information for word tokens

## Identification of Negative Expressions

Words like adjectives and verbs can convey the opposite feeling simply by adding a neg- ative prefix to the beginning of the word. Consider the following statement, which was unearthed in the course of an analysis of an electrical product: "The built-in speaker also has its purposes but so far nothing new." The word "revolutionary" is included in the list of suitable words, as the list indicates. On the other hand, the term "nothing ground- breaking" almost always evokes feelings of disappointment. As a direct consequence of this, it is absolutely necessary to understand these terms. Both the negation-of-adjective (NOA) and the negation-of-verb phrase types have been uncovered over the course of this research (NOV).

The POS tagger treats the most prevalent negative prefixes, such as not, no, and nothing, as adverbs. As a result, we suggest a particular algorithm for phrase identi- fication. The system was able to recognize 21,586 distinct phrases, each of which includes a negative prefix, with a total incidence of over 0.68 million.

## The real-world labels

The technique for classifying sentiment polarity consists of two steps: the first is the classification of sentences, and the second is the categorization of reviews. The purpose of classifying sentences at the sentence level is to determine if a given sentence expresses a good or negative feeling depending on the mood it expresses. It is necessary to provide ground truth tags for the training data for this classification strategy. These tags should specify whether a particular statement is positive or negative. However, due to the sheer amount of data that we possess, the task of ground truth classification has become an extremely challenging obstacle. An approach that uses machine tagging is utilized as a solution since it is physically impossible to tag each sentence individually by hand. The tactic makes use of a bag-of-words model, which just counts how many times positive or

20 *CHAPTER 6. METHODOLOGY AND RESEARCH DESIGN*

negative (word) tokens appear in each sentence. If there are a greater number of positive tokens than there are negative tokens, then the phrase will be labeled as positive, and vice versa. Similar approaches were taken in the tagging of the Sentiment 140 Tweet Corpus, both of which are analogous to this one. For the purpose of review-level classification, the ground truth tags, also known as star-scaled ratings, are already incorporated into the training data.

## Formation of Feature Vectors

Sentiment tokens and scores are generated based on information gathered from the orig- inal dataset. They are used to classify emotions and are sometimes referred to as char- acteristics. In order to train the classifiers, it is necessary to transform every piece of the training data into a feature vector, a vector comprising those features. In order to classify at the sentence level (review level), a feature vector is generated from a sentence (review). The dimensional control of each vector presents a challenge.

Because of the curse of dimensionality, classifiers need that all vectors have the same amount of dimensions, and a vector shouldn’t have thousands or hundreds of features or feature values. This is a problem that only applies to emotional gifts: On the one hand, there are 11,478-word tokens and 3,023 phrase tokens, but the vectors cannot be produced by simply adding the tokens that were present in a sentence (or review), as sentences and reviews contain varying amounts of tokens. We get around this by using two binary strings to signify the existence of each emotion token within a sentence or review, as we are only interested in how these tokens are used within those contexts. A string of 11,478 bits is used for word tokens, whereas a string of 3,023 bits is used for phrase tokens. In the end, instead of putting the inverted texts directly into a feature vector, a hash value is calculated using Python’s built-in hash function and stored. Thus, a sentence-level feature vector consists of a ground truth label, an average emotion score, and two hash values generated from the flipped binary strings.

*6.5. FORMATION OF FEATURE VECTORS* 21

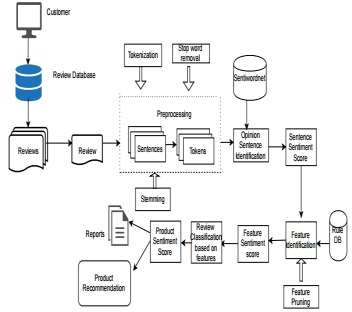


Figure 6.1: System Diagram

# Technologies That Are In Use

## Python

Python is a popular dynamic programming language that has a straightforward syntax and an indentation structure that is simple to understand.

## Stanford CoreNLP

It is an application for processing natural language. It has a comprehensive toolkit with a good range of grammar-checking tools. It is both quick and dependable. It does this by determining the grammatical function of each word in the phrase. It may be bent and stretched as needed.

## Beautiful Soup

Beautiful Soup is a scraping package written in Python that was developed by a third party called Crummy.

22

* 1. *SCIKIT-LEARN* 23

## Scikit-learn

Python’s Scikit-learn is a machine learning toolkit that includes implementations of a wide variety of machine learning techniques. These machine-learning methods include classification, regression, clustering, and others.

# Sentiment Classification Algorithms

Throughout the course of our investigation, we made use of scikit-learn, which is a Python- based machine learning software tool that was built by the developers at MIT. In order to classify the information, we will utilize three different models: a Naive Bayesian model, a Random Forest model, and a Support Vector Machine model. These models were decided upon as the ones to use for the categorization procedures.

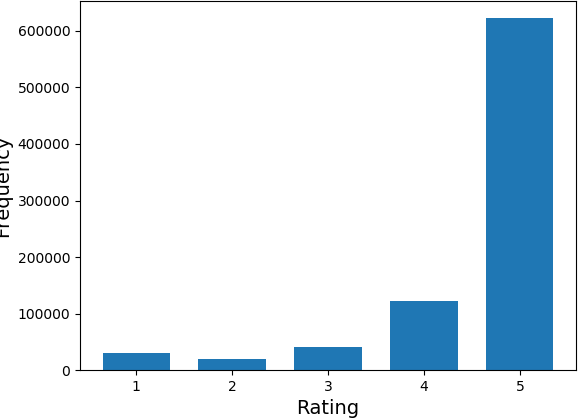


Figure 8.1: Bar Chart showing Frequency of Ratings 24

* 1. *LINEAR REGRESSION* 25

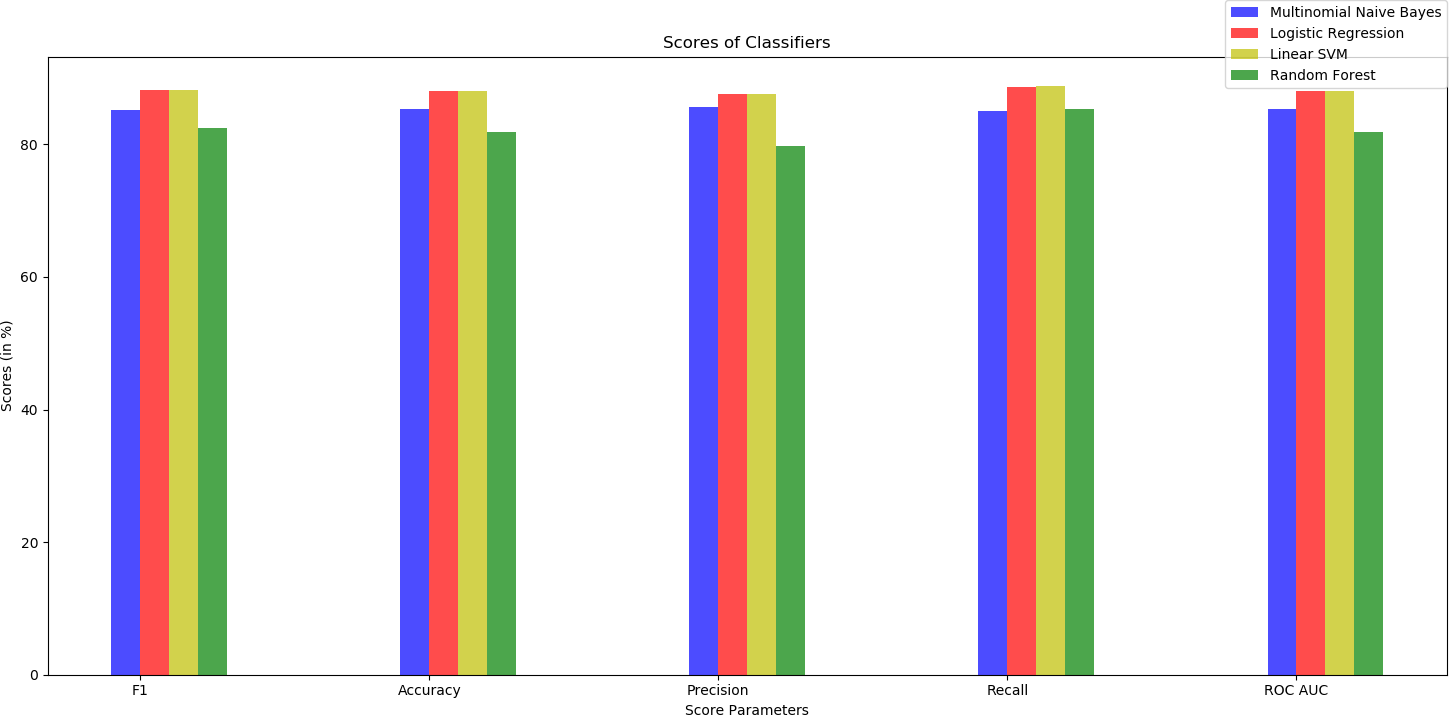


Figure 8.2: Bar Chart for Accuracy, Precision, Recall, and Roc-Auc

The subject of Artificial Intelligence (AI) known as "mechanical learning" in- vestigates the ways in which machines may learn new knowledge and abilities, as well as the ways in which they can identify information that they have already learned. Data mining, computer vision, the processing of natural languages, search engines, biometrics, medical diagnostics, credit card fraud detection, a market analysis of stocks, DNA se- quence, speech and handwriting recognition, robotics, and strategy games are just some of the fields that have found widespread use for machine learning. The following are some of the most often-used algorithms for machine learning:

## Linear Regression

The value of the dependent or reliant variable is estimated using independent variables for statistical procedures in linear regression, which is stated as the definition of linear regression. The process of mapping a dependent variable and an independent variable onto a line is known as the regression line.

26 *CHAPTER 8. SENTIMENT CLASSIFICATION ALGORITHMS*

## Logistic Regression

This strategy is used to determine the discrete dependent variable from the set of different variables. Logistic regression is a type of statistical analysis. The coefficients necessary to estimate a probability logistic transformation can be obtained by logistic regression.

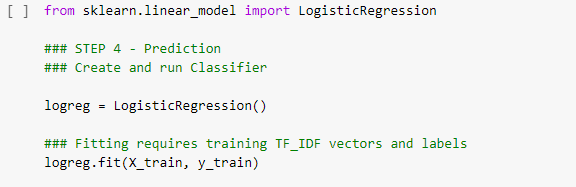


Figure 8.3: Logical Regression Classifier Model

## Decision Tree

Decision tree is a tool that may be used for classification as well as regression. It has a structure that looks like a tree. The most useful characteristic of a dataset is entered into an algorithm for generating decision trees; following this step, the training data set is segmented into subsets. In order to develop a training model that can accurately predict the class or value of the destination variable, decision trees are constructed.

* 1. *SUPPORT VECTOR MACHINE (SVM)* 27

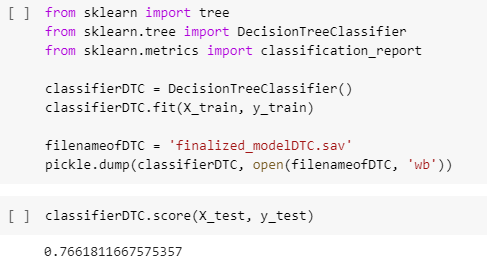


Figure 8.4: Decision Tree Classifier Model



Figure 8.5: Decision Tree Algorithm

## Support Vector Machine (SVM)

A Binary Classifier (BC) is an example of an SVM (SVM). The row data is drawn on the point that is n-dimensional. A hyperplane that divides the data sets is drawn as part of this process. This improved separation maximizes the margin of error for the training

28 *CHAPTER 8. SENTIMENT CLASSIFICATION ALGORITHMS*

data.

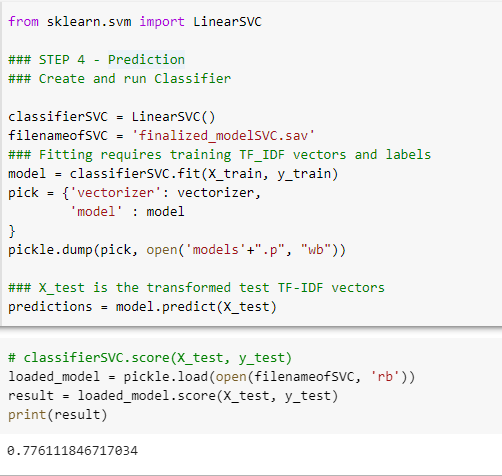


Figure 8.6: Linear SVM Model

* 1. *NAIVE BAYES* 29

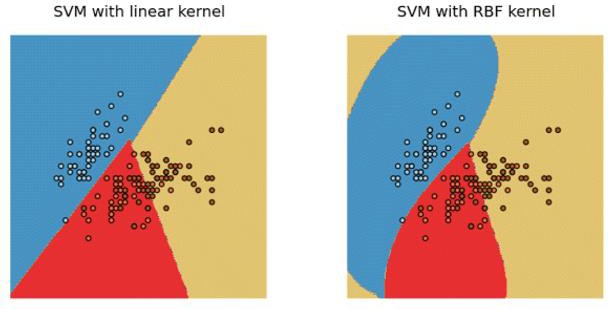


Figure 8.7: SVM Mapping

## Naive Bayes

This method is based on the theory of Bayes, which is used by more advanced classification methods. Naive Bayes is a classification method. The method in question is one of categorization. It learns how it is possible for an entity to belong to a specific category or class if it possesses a certain set of qualities.



Figure 8.8: Naive Bayes Classifier Equation

30 *CHAPTER 8. SENTIMENT CLASSIFICATION ALGORITHMS*

## KNN

KNN is a classification and regression algorithm that utilizes this method. This is a straightforward example of an algorithm for machine learning. It makes a copy of the cases and then searches for new information in the k-neighborhoods that are most similar to it. It is a lifesaver for the cases. When provided with a dataset for testing, KNN generates accurate predictions.

## K-means Clustering

K-means Clustering is a technique of unsupervised that facilitates reaching the highest potential cap that may be achieved. The initial phase in the partitioning method was using the Euclidean distance to organize the datasets into clusters so they could be more easily analyzed.

## Random Forest

Random Forest belongs to the class of algorithms known as supervised computations. A technique referred to as a random forest generates a set of numerous classification trees by combining the results of a number of separate decision trees. Both regression and classification can benefit from using this information. Rules are provided by the Decision Tree algorithm for the given training data set, which includes both targets and features.

*8.9. DIMENSIONALITY REDUCTION ALGORITHMS* 31

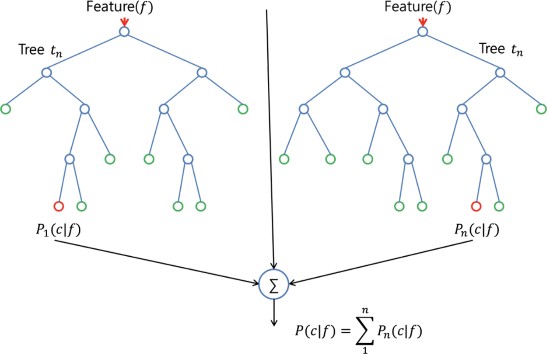


Figure 8.9: Random Forest Tree

## Dimensionality Reduction Algorithms

This indicates that the number of random variables can be minimized by getting those crucial variables. The reduction of dimensionality can be accomplished through the ex- traction of functions and the selection of features. Principal Component Analysis, ab- breviated as PCA, is a technique for extracting major variables from a large number of variables. This technique can be used to perform the main component analysis.

32 *CHAPTER 8. SENTIMENT CLASSIFICATION ALGORITHMS*

## Gradient Boosts and Ada Boost

The classification and regression algorithms are known as gradient boosts and ada boosts The algorithm of gradient boosts is known as the classification and regression algorithm. Only the features that improve model prediction are considered for selection by AdaBoost. In order for it to function, it first chooses a fundamental algorithm, such as decision trees, and then iteratively improves that algorithm by taking into consideration incorrect examples that are present in the training data set.

# Results and Discussions

## Classification of Multiclass

For the sake of this analysis, Amazon customer evaluations of mobile phones were divided into three distinct groups: positive, negative, and neutral, based on each product’s overall star rating. Reviews that had one or two stars were regarded as being unfavorable, reviews that had four or five stars were regarded as being extremely favorable, and reviews that had three stars were regarded as being neutral. As a direct result of this, every one of the proposed models was put through its paces by employing every single method of feature extraction.

In addition to this, the results of the evaluation of the RF model employing a variety of different feature extraction methodologies were carried out. The RF with Glove has an accuracy that is 90 percent better than earlier approaches, while suffering just a 39 percent loss in cross-entropy. This is in comparison to other methods.

The assessment results of the NB classifier using each of the suggested feature extraction techniques in this research, with the exception of the Glove approach, are also represented in the figure. Although the NB theory assumes that features are un- connected to one another, the glove is a word vector representation that groups together words that are semantically similar to one another. When compared to the other ways, the NB classifier with BOW (Trigram) has a higher level of accuracy than the other ap- proaches; however, it also has a greater level of cross-entropy loss, with 78 percent and 1.18, respectively, when contrasted with the other methods.

if we take into account the results of running Bi-LSTM with two separate em- beddings, namely, the fine-tuned Glove embedding and the jointly trained embedding, we find that the results are as follows: Both of these methods were utilized by the model in order to achieve the desired levels of accuracy (93 percent) and cross-entropy loss (0.189). Because of the way the Bi-RNN is organized, it is feasible for networks to be trained in both temporal directions at the same time. This is made possible by the structure of the Bi-RNN. The outcome was better than expected (backward and forward). Because of this structure, it is feasible to include information from both the beginning and the conclusion of a series, which, in the end, leads to the increased performance of the model.

In addition to this, the so-called "confusion matrix" was applied in order to describe the performance of categorization models. The BERT model performed signif- icantly better than the alternatives to classification that were considered. On the other hand, as can be seen from the figure, the BERT model’s success in classifying the neutral

33

34 *CHAPTER 9. RESULTS AND DISCUSSIONS*

class achieved only 65 percent; and this was also something that was anticipated, given that a review with a rating of three stars does not necessarily mean that the customer was completely balanced in their opinion between positive and negative aspects of the product or service. This was also something that was anticipated, given that a review with a rating of three stars does not necessarily mean that the customer was completely balanced in their opinion between positive and negative In contrast, it had a classification score of 97 percent for reviews that were positive and a classification score of 94 percent for reviews that were critical.

## Binary classification

After putting each suggested model through its paces in multiclass classification via a broad variety of feature extraction approaches, we next analyzed the results of our ex- periment. In this experiment, Amazon mobile phone reviews were divided into positive and negative groups according to the star rating. Ratings of "one star" and "two stars" are viewed as being negative, whilst ratings of "four stars" and "five stars" are seen as being favorable. As a direct result of this fact, the binary classification will be carried out making use of the same feature extraction procedures that were put into play in order to arrive at the conclusive results of the multiclass classification.

When looking at the results of binary classification for all models on the test dataset, it is evident that BERT delivered a superb result with the best accuracy of 98 percent. When looking at these findings, it is important to note that BERT had the highest accuracy. This is the situation that has arisen due to the fact that BERT was the model that attained the highest percentage. Furthermore, the results of the Bi-LSTM suggested a degree of accuracy that was comparable to 97 percent of the time. The accuracy of the baseline model of RF with Glove was, on the other hand, better to that of both the LR and the NB. This strategy was successful in earning a score of 90 percent. The BERT model, in comparison to the designs of other goods, offers the best level of accuracy in binary categorization. Once recasting the job as a binary-classification issue, we discovered that the overall performance of every model was greatly improved after the recasting was complete. As a result, the performance of the model is in some manner influenced by the neutral class.

* 1. *PERFORMING AN EFFECTIVENESS ANALYSIS ON THE BERT MODEL* 35

## Performing an effectiveness analysis on the BERT model

Since it has a grade of 94 percent for its correctness, the BERT model is the one that is considered to be the most accurate one currently available. As a direct result of this, we will need to conduct a performance evaluation in order to figure out how precisely its feelings can be grouped together. A significant amount of focus is presently being directed into the interpretability of models as well as the investigation of potential faults in categorization. It is necessary for us to have a knowledge of the characteristics that cause some models to provide inaccurate classifications in particular reviews while generating appropriate identifications in others. As a consequence of this, we searched for the reviews that had the highest loss values and the greatest number of incorrect classifications. Because we believed that they would describe the decision boundary of the BERT model in the same manner that they did, we read them as describing it in that manner. It was unequivocally established that the BERT model assigned a probability of 0.99 to the verdict of "positive," indicating that the evaluation was positive. This was evidenced by the fact that the evaluation was positive.

Despite the fact that the evaluation did not indicate whether the item was favorable or unfavorable. This is an example of a circumstance in which the BERT model is true but the label is not right. The model recognizes the existence of the positive word "love" and draws attention to it by coloring it in a shade of green whose intensity changes in relation to that of other positive phrases. The customer gave the products just three out of a possible five stars, despite the fact that he had previously said that he was completely happy with the purchase. Because of this, it was noted at the beginning of the text that ratings of three stars do not necessarily indicate that a person has an objective judgment. In addition to this, it has also been seen that the customer has a negative attitude toward the mobile phone in the way that he expresses his viewpoint. This was discovered by observing the way in which he communicates his opinion. In point of fact, the model accurately classifies the review as being negative, despite the fact that the actual class was wrongly given as positive.

A random sample approach is utilized in the conduct of the experiment, which is carried out using a dataset that contains 200,000 customer reviews. In order to find a solution to the problem of class imbalance, a total of 40,000 reviews from each of the five rating classes are taken into consideration.

Following this, the reviews are analyzed and preprocessed using the methodology that we have proposed. Eighty percent of the reviews are used for training, while the remaining twenty percent are put toward testing the models. After then, five distinct categorical classes are generated for ordinal values of overall ratings ranging from one

36 *CHAPTER 9. RESULTS AND DISCUSSIONS*

Name of classifier Multinomial

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| F1 | Accuracy | Precision | Recall | ROC AUC |
| 85.25% | 85.31% | 85.56% | 84.95% | 85.31% |
| 88.12% | 88.05% | 87.54% | 88.72% | 88.05% |
| 88.12% | 88.11% | 87.59% | 88.80% | 88.11% |
| 82.43% | 81.82% | 79.74% | 85.30% | 81.83% |

NB

Logistic Regression Linear SVC

Random Forest

Table 9.1: Table of Accuracy, Precision, Recall and Roc Auc

to five. This is done so that the rating prediction can be approached as a multi-class classification problem. There are three different multi-class classifier models that are put to use nowadays. These models are the Multinomial Naive Bayes Classifier, the Logistic Regression Classifier, and the Linear SVC Classifier (Support Vector Classifier).

These models are trained and used to classify reviews into one of the five classes by utilizing TF-IDF vector features obtained from review text in addition to other derived characteristics such as polarity and the length of the review. These features are used in conjunction with the review text. Below is a description of both the outcomes that were obtained and the measures that were used for the evaluation.

## Confusion Matrix and Accuracy

The "confusion matrix" is an essential criterion that is utilized in the process of deter- mining the efficacy of classifiers. It is the number of samples that have been successfully categorized, and it is used in the computation of precision, recall, and F1 scores. All of the classifiers predict reviews with 1 star and 5 stars with a great deal more accuracy than the neutral reviews, which is something that can be seen quite plainly (3 stars). Any random prediction would only be correct 20 percent of the time due to the fact that we have five rating classes; this percentage serves as our baseline. It can be seen that all of the classifiers perform significantly better than the baseline, but the Logistic The regression classifier achieves the highest level of success, with an overall accuracy of

54.1 percent and an increase of almost 170 percent over the performance of the random classifier.

*9.5. A PLOT SHOWING THE RECEIVER’S OPERATING CHARACTERISTICS* 37

## A plot showing the receiver’s operating char- acteristics

The Receiver Operating Characteristic, also known as the ROC, is a probability curve, and the area under the curve, also known as the AUC, is a measure of separability. This evaluates the model’s ability to differentiate across different types of data. For multi- class classification models, the ROC curve is an important metric that is more important than accuracy because it visualizes the model’s accuracy across all conceivable thresholds. A ROC area that has a value of 0.5 is considered to be the baseline for a model that represents random classification.

It is possible to see that the AUCs are significantly higher than the baseline, and this is especially true for the ROC curves that represent ratings of one star and five stars. It shows that these two classes are significantly more distinguishable from one another and can be easily categorized, but class 3, which represents neutral evaluations with a rating of three stars, is the class that is the least distinguishable.

Because of this Training of Public reviews dataset, the machine is now able to determine whether a given statement will produce a favorable reaction or a negative one based on the responses it has previously received. Precision, also referred to as just a positive predictive value, is the fraction of relevant instances that are included in the instances that have been retrieved, whereas recall, which is also known as sensitivity, is the fraction of relevant instances that have been retrieved in relation to the total number of relevant instances. Both of these concepts are related to the concept of sensitivity. Precision and recall are two different ways of measuring accuracy. Because of this, recall and precision are dependent on having a grasp of, and being able to quantify relevance. A test’s degree of precision may be evaluated using a metric called the F1 score, which is sometimes referred to as the F score or the F measure. When determining the final score, it takes into consideration both the accuracy p and the recall r of the examination. The recall is calculated by dividing the number of accurate positive results by the total number of relevant samples. Precision is calculated by dividing the number of correct positive results by the total number of positive results produced by the classifier.

The F1 score is calculated by taking the harmonic average of the precision and recall scores, with a score of 1 representing flawless precision and recall and a score of 0 representing the worst possible performance. A receiver operating characteristic curve, also known as a ROC curve, is a graphical plot that displays the diagnostic capabilities of a binary classifier system when its discrimination threshold is modified. This plot is used in the field of statistics. The idea behind the Relative Operating Characteristic, also known as ROC, is built upon the Total Operating Characteristic, also known as TOC. TOC displays all of the information that is contained in the two-by-two contingency table

38 *CHAPTER 9. RESULTS AND DISCUSSIONS*

for each threshold. Because the ROC only supplies two pieces of relative information for each threshold, the TOC offers a great deal more information than the ROC does. When using normalized units, the area under the curve, which is frequently referred to as simply the AUC, is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (presuming that ’positive’ ranks higher than ’negative’). This is the case regardless of whether or not ’positive’ ranks higher than ’negative’. This probability is equivalent to the likelihood that a classifier would place a randomly selected positive instance in a higher ranking than a randomly selected negative instance. The following is an example of one interpretation that may be given for this: For the purpose of calculating the area under the curve, the following equation can be used: (the integral boundaries are inverted since a large T results in a smaller value on the x-axis). In order for the computer to evaluate how well the training was carried out, it takes into account not only how accurately the data was recalled but also F1.

Calculations are being made on the evaluation matrix of confusion.

As a result, it is able to determine whether a review written by an outside party is positive or negative.

A review that is seen to be good will be denoted by the number [1], while a review that is deemed to be negative will be denoted by the number [-1] and a neutral review would give out [0].

## Methods of Evaluation

The effectiveness of any classification model is judged according to the model’s average F1-score, which is comprised of the following factors:

where P I represents the precision of the ith class, R I represents the recall of the ith class, and n is the total number of classes. P I and R I are assessed by the utilization of 10-fold cross-validation. The following constitutes an application of 10-fold cross-validation: A dataset is divided into 10 subsets of equal size, each of which contains 10 positive class vectors and 10 negative class vectors. The total number of class vectors in the dataset is 100. One of the ten subsets is utilized as the validation data for testing the

*9.7. RESULT ON MANUALLY-LABELED SENTENCES* 39

classification model, while the other nine subsets are put to use as training data. There are a total of 10 subsets. After that, the procedure of cross-validation is carried out ten times, with each of the ten subsets being utilized as the validation data precisely once. After that, we use the average of the ten different folds’ results to get a single estimation. ROC (Receiver Operating Characteristic) curves are presented when training data are labeled under two classes (positive and negative) for sentence-level categorization. This allows for a more accurate comparison of the sentences’ levels of performance.

## Result on manually-labeled sentences

The carefully annotated sentences serve as the basis for the generation of 200 feature vectors. As a consequence of this, the performance of the categorization models is com- parable based on their F1 scores, with all three scores having the same value of 0.85 in accordance with this comparison. It is easy to see, with the assistance of the ROC curves, that all three models did pretty well when evaluating data that have a high poste- rior probability. (A posterior probability of a testing data point, denoted by the symbol P(+|A), is calculated by the classification model as an estimate of the chance that A will be categorized as positive.) The Naive Bayes classifier performs better than the SVM classifier, with a greater area under curve, as the probability decreases. The Random Forest model achieves the highest levels of accuracy, on average.

## Result on machine-labeled sentences

Two million feature vectors, one million of which have positive labels and the other million having negative labels, are constructed from two million machine-labeled phrases, which are together referred to as the whole set. The whole set is divided into four subsets, with subset A containing 200 vectors, subset B containing 2,000 vectors, subset C containing

40 *CHAPTER 9. RESULTS AND DISCUSSIONS*

20,000 vectors, and subset D containing 200,000 vectors, respectively. Subset D comprises all of the vectors in the complete set. In each and every one of the subsets, the number of vectors that have positive labels and the number of vectors that have negative labels is precisely equivalent to one another. Following that, the effectiveness of the classification models is assessed based on a total of five separate vector sets (four subsets and one complete set).

The F1 scores of the models are consistently improving as more training data is being added to the models. The SVM model experiences the most dramatic improvement, going from 0.61 to 0.94 as a result of an increase in the number of training data from 180 to 1.8 million. This model outperforms the Naive Bayes model and rises to the position of second best classifier, both on subset C and over the entire data set. Once again, the Random Forest model achieves the best results for datasets that encompass all scopes.

## Review-level categorization

For the purpose of categorization, three million feature vectors are produced. The vectors that are formed from reviews with ratings of at least four stars are denoted as positive, whilst the vectors that are generated from reviews with ratings of one star or two stars are denoted as negative. The preparation of neutral class vectors involves using reviews with a rating of three stars. As a direct consequence of this, the total collection of vectors has been neatly classified into one of three categories: positive, neutral, or negative. In addition, three subsets are derived from the whole set. Subset A contains 300 vectors, subset B contains 3,000 vectors, subset C contains 30,000 vectors, and subset D contains 300,000 vectors, respectively. Subset D comprises all of the vectors in the complete set.

In terms of how well they function, the SVM model and the Naive Bayesain model appear to be very identical. This is something that can be seen without a doubt. On every vector set, both of these models perform significantly better than the Random Forest model. However, when the models are employed to classify sentences at the sen- tence level, neither of them can reach the same level of performance as the other because of their relatively poor results on the neutral class.

The experimental outcome shows promise, both in terms of sentence-level cat- egorization and review-level categorization. Specifically, sentence-level categorization is

*9.9. REVIEW-LEVEL CATEGORIZATION* 41

|  |  |
| --- | --- |
| 68556 | 11470 |
| 12032 | 67942 |

Table 9.2: Confusion Matrix for Multinomial NB

|  |  |
| --- | --- |
| 69928 | 10098 |
| 9023 | 70951 |

Table 9.3: Confusion Matrix for Logistic Regression

encouraging. It was found that the averaged sentiment score is a powerful feature on its own, as it is able to produce an F1 score of more than 0.8 for the sentence-level cate- gorization when applied to the full set of data. This was one of the observations made. The feature has the potential to generate an F1 score that is higher than 0.73 when used in conjunction with the full set for the review-level categorization. Having said that, there are still a handful of restrictions that come with this study. The first issue is that if we wish to categorize reviews according to the precise star-scaled ratings they have, it makes the process of categorizing reviews at the review level more complicated. In other words, the F1 scores that are derived from these kinds of tests have values that are significantly lower than 0.5. The second limitation is that our scheme for analyzing sen- timents, which was proposed in this study, is dependent on the occurrence of sentiment tokens. Therefore, it is possible that our scheme will not work well for those reviews that only contain implicit sentiments because our scheme relies on the occurrence of sen- timent tokens. Because an implicit sentiment is typically communicated through some neutral phrases, determining the polarity of the sentiment it conveys can be challenging. For instance, a line like "Item as described.", which appears quite frequently in positive reviews, is made up entirely of terms that are not positive or negative.

Keeping these constraints in mind, our work going forward will concentrate on finding solutions to the problems that have been raised. To be more specific, additional features will be extracted and arranged into feature vectors so that review-level catego- rizations can be improved. When it comes to the problem of analyzing implicit sentiment, the next thing we need to do is figure out how to identify the presence of a given sen- timent within the context of a specific product. In an additional upcoming study, our categorization approach will be tested using datasets from various sources.

|  |  |
| --- | --- |
| 69963 | 10063 |
| 8955 | 17019 |

Table 9.4: Confusion Matrix for Liner SVC

# Contribution of Individual Team Member

The table below displays the estimated percentage contributions made by each team member to this project:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Requirement Analysis | Planning | System Analysis and Design | Implementation |
| Sadman | 50 | 50 | 70 | 30 |
| Maruf | 50 | 50 | 30 | 70 |
| **Total** | 100 | 100 | 100 | 100 |

Table 10.1: Contribution of Individual Team Member

42

# Bibliography

1. A Study on Sentiment Analysis: Methods and Tools. *International Journal of Sci- ence and Research (IJSR)*, 4(12):287–292, dec 5 2015.
2. Sentiment Classification using Machine Learning Techniques. *International Journal of Science and Research (IJSR)*, 5(4):819–821, apr 5 2016.
3. Emotion Detection Using Machine Learning. *International Journal of Recent Trends in Engineering and Research*, 3(6):28–32, jun 7 2017.
4. Sentiment Analysis for Social Media using SVM Classifier of Machine Learning. *In- ternational Journal of Innovative Technology and Exploring Engineering*, 8(9S4):39– 47, oct 1 2019.
5. Product RECOMMENDATION SYSTEM USING MACHINE LEARNING. *Inter-*

*national Research Journal of Modernization in Engineering Technology and Science*, dec 10 2022.

1. M. Birjali, A. Beni-Hssane, and M. Erritali. Machine Learning and Semantic Senti- ment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks. *Procedia Computer Science*, 113:65–72, 2017.
2. S. D. and P. M. Enhanced Sentiment Analysis by using Social Media. *International Journal of Computer Applications*, 182(15):27–30, sep 17 2018.
3. F. Dalipi, K. Zdravkova, and F. Ahlgren. Sentiment Analysis of Students’ Feedback in MOOCs: A Systematic Literature Review. *Frontiers in Artificial Intelligence*, 4, sep 9 2021.
4. M. Ghosh and G. Sanyal. An ensemble approach to stabilize the features for multi- domain sentiment analysis using supervised machine learning. *Journal of Big Data*, 5(1), nov 14 2018.
5. T. Hariguna, W. M. Baihaqi, and A. Nurwanti. Sentiment Analysis of Product Re- 43

44 *BIBLIOGRAPHY*

views as A Customer Recommendation Using the Naive Bayes Classifier Algorithm. *IJIIS: International Journal of Informatics and Information Systems*, 2(2):48–55, sep 1 2019.

1. J. Hausler, J. Ruscheinsky, and M. Lang. News-based sentiment analysis in real estate: a machine learning approach. *Journal of Property Research*, 35(4):344–371, oct 2 2018.
2. N. Jhmat. Sentiment-based Movie Recommender System using Deep Learning. *Pak- istan Social Sciences Review*, 2:295–304, jun 30 2018.
3. G. Kaur. Text Mining Based Approach to Customer Sentiment Analysis Using Ma- chine Learning. *Journal of Advances and Scholarly Researches in Allied Education*, 15(6):58–65, jul 1 2018.
4. R. Kaur. Sentiment Analysis of Movie Reviews: A Study of Machine Learning Algo- rithms with Various Feature Selection Methods. *International Journal of Computer Sciences and Engineering*, 5(9), 9 2017.
5. N. Kewsuwun and S. Kajornkasirat. A sentiment analysis model of Agritech startup on Facebook comments using naive Bayes classifier. *International Journal of Elec- trical and Computer Engineering (IJECE)*, 12(3):2829, jun 1 2022.
6. A. Kumar and T. M. Sebastian. Sentiment Analysis: A Perspective on its Past, Present and Future. *International Journal of Intelligent Systems and Applications*, 4(10):1–14, sep 1 2012.
7. S. Kumar. Using Machine Learning and Sentiment Analysis in Official Statistics.

*SSRN Electronic Journal*, 2018.

1. B. M. and V. B. Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis. *International Journal of Computer Applications*, 146(13):26–30, jul 15 2016.
2. W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and appli- cations: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 12 2014.

*BIBLIOGRAPHY* 45

1. V. Naramula and K. A. Sentiment analysis in aspect term extraction for mobile phone tweets using machine learning techniques. *International Journal of Pervasive Computing and Communications*, oct 18 2021.
2. P. Poomka, N. Kerdprasop, and K. Kerdprasop. Machine Learning Versus Deep Learning Performances on the Sentiment Analysis of Product Reviews. *International Journal of Machine Learning and Computing*, 11(2):103–109, 3 2021.
3. H. Priyanka and R. Ashok Kumar. Sentiment Analysis using Machine Learning Based Ensemble Model for Food Reviews. *International Journal of Innovative Re- search in Applied Sciences and Engineering*, 4(3):690–694, sep 1 2020.
4. M. Rokaya. Sentiment Analysis of Social Media Micro Blogs using Power Links and Genetic Algorithms. *International Journal of Computer Applications*, 177(44):28–35, mar 17 2020.
5. S. S. and P. K.V. Sentiment analysis of malayalam tweets using machine learning techniques. *ICT Express*, 6(4):300–305, 12 2020.
6. J. Singh, G. Singh, and R. Singh. Optimization of sentiment analysis using machine learning classifiers. *Human-centric Computing and Information Sciences*, 7(1), 12 2017.
7. N. K. Singh and D. Singh. Feature Fusion for Negation Scope Detection in Senti- ment Analysis: Comprehensive Analysis over Social Media. *International Journal of Advanced Computer Science and Applications*, 10(5), 2019.
8. S. P. Singh, S. Karkare, S. M. Baswan, and V. P. Singh. Unsupervised Machine Learning for Co/Multimorbidity Analysis. *International Journal of Statistics and Probability*, 7(6):23, sep 12 2018.
9. A. Tripathy and S. K. Rath. Classification of Sentiment of Reviews using Super- vised Machine Learning Techniques. *International Journal of Rough Sets and Data Analysis*, 4(1):56–74, 1 2017.
10. C. C. Yang, X. Tang, Y. C. Wong, and C. P. Wei. Understanding Online Consumer

46 *BIBLIOGRAPHY*

Review Opinions with Sentiment Analysis using Machine Learning. *Pacific Asia Journal of the Association for Information Systems*, pages 73–89, 2010.

1. S. Yi and X. Liu. Machine learning based customer sentiment analysis for recom- mending shoppers, shops based on customers’ review. *Complex Intelligent Systems*, 6(3):621–634, jun 11 2020.

[16] [19] [25] [29] [30] [1] [3] [7] [5] [11] [13] [6] [9] [10] [12] [15] [21] [23] [27] [28] [18] [14]

[17] [4] [22] [26] [8] [24] [20] [2]