National College of Ireland

Sentiment Analysis Techniques for Restaurant Reviews Across Multiple Attributes

MSc Research Project
MSc Data Analytics

Pawan Kumar
Student ID: x22186115

School of Computing
National College of Ireland

Supervisor: Dr. Muslim Jameel Syed

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Pawan Kumar		
Student ID:	X22186115		
Programme:	MSc Data Analytics	**Year:**	2023-24
Module:	MSc Research Project		
Supervisor:	Dr. Muslim Jameel Syed		
Submission Due Date:	12 August 2024		
Project Title:	Sentiment Analysis Techniques for Restaurant Reviews Across Multiple Attributes		
Word Count:	7678		
Page Count	**22**		

I hereby certify that the information contained in this (my submission) is information pertaining to Research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Pawan Kumar
Date:	12 August 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□

Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Sentiment Analysis Techniques for Restaurant Reviews Across Multiple Attributes

Pawan Kumar

X22186115

Abstract

The spread of the Internet and social media have made the world more connected and nearer, where people share their opinions, ideas, and pictures with the public. Exploiting this information has become the most popular for companies around the world, and sentiment analysis is one technique out of it. The online restaurant business has made sentiment analysis a critical tool for understanding consumer opinions and improving decision-making. The study focuses on implementing sentiment analysis on the review data to extract the opinion of customers as positive or negative using a standard methodology, KDT (Knowledge Discovery in Text). Various machine learning models like Logistic Regression, Xg Boost, Random Forest and Naïve Bayes and deep learning models (LSTM)are used in research focusing on extracting useful information from text data using Natural Language Understanding (NLU) and NLP (Natural Language Processing). To convert the text data into numbers, metrics like TF-IDF, Count Vectorizer and Word2vec are used where the LSTM model with Count Vectorizer attained the maximum accuracy of 95.4%. The research has showcased a comprehensive analysis of various techniques tailored for the restaurant industry, offering valuable information for academic as well as business applications.

1 Introduction

1.1 Background

The spread of the Internet around the world has led to an increase in online activities like social media, online surfing, and E-commerce. The following usage has created a lot of opportunities around the world for various businesses. Customers are interacting with brands in terms of product satisfaction, and a new trend has highlighted the concept of reviews on these sites. These reviews are used as a source of information by users before buying any product or service. According to a survey, 90% of online shoppers tend to read a survey before visiting any business. (*Local Consumer Review Survey 2024*, n.d.)

Moreover, any negative review prevents them from visiting any business. The social world tends to confirm a purchase if it is approved by the majority of people, and with these reviews, one can understand the product or services in a better way. The given topic is affecting the restaurant business, where positive reviews tend to increase the loyalty and the footfall of a restaurant, and negative reviews decrease the reputation. These reviews work as mouth-to-mouth advertising for restaurants and let customers try new foods and outlets. Restaurants should track these reviews for better online reach.

1.2 Motivation

Customer satisfaction is the main goal of any business or service. These online reviews can also do the same in terms of E-commerce or restaurant business. With the information available in their hands, people tend to search on Google or look out at reviews before deciding where to eat. By analyzing these reviews, the restaurant can get a deeper understanding of the customers' opinions and use the information provided for better offerings and customer experience. Dublin offers a variety of cuisines because people from multiple demographics live there. From Wollen Mills to fine dining like Chapter One, Taza, Indian cuisine, and others, the city offers wide variety, and understanding the people's tastes and preferences is one issue. Sentiment analysis is one approach that is used to find the sentiment behind the reviews, whether it is positive, negative, or neutral. The technique is backed up by AI-based algorithms that can learn from the text data using NLP (Natural Language Processing). The approach helps to get crystal clear signals about the services and the food. The given approach helps the restaurant find opportunities to improve and provide a personalized experience. These reviews help the restaurant to maintain an online image and optimize its operations. In the end, every business wants higher profits and a better reputation, and the given feedback management helps the business, especially the restaurant, to enhance the dining experience and reduce customer churn.

1.3 Research Objectives

1. Develop various machine learning and deep learning frameworks to identify the best algorithm for the classification of the sentiment of restaurant reviews.
2. Understand the customer opinion around various aspects of restaurant experience from cuisines, services, ambience and satisfaction and build a pipeline to understand the sentiment based on that.
3. Provide insights to the restaurant, enabling them to focus on areas of improvement, customer satisfaction, and data-driven decisions.
4. Understand the domain of natural language processing and build machine learning and deep learning architectures to extract meaningful patterns for text data for a given domain.

1.4 Research Questions

1. Understand the effectiveness of machine learning as well as deep learning architectures in classifying restaurant reviews.
2. Explain the text data and the keywords used and improve the decision-making process, focusing more on the implementation of explainable models.
3. What is the effect of text metrics and the type of model chosen for the given dataset?

1.5 Limitations

1. Natural language showcases ambiguity based on the context of data and may lead to inappropriate results while building these models.

2. Restaurant reviews data share a general aspect and provide information about the cuisine, food, and services. Hence, identifying these themes is challenging with respect to sentiment analysis.
3. Data containing emojis, slang, and other noise can reduce the accuracy of these models.
4. Machine learning models are generally unable to capture the context and meaning of the words.
5. Customer reviews generally contain personal experiences and may lead to bias in the modelling.

1.6 Contribution of the Research

The given solution provides a detailed analysis of the text data using Natural Language Understanding (NLU) to generate models that can correctly classify the reviews based on the sentiment and provide insights to the restaurant owners.
The project highlights the use of KDT (Knowledge Discovery In Text), which is the extension of Knowledge Discovery in Databases (KDD).(Sukanya & Biruntha, 2012). The following technique is a combination of analyzing text data to find the summaries, patterns, and trends to analyze the text data. The main aim of the research is to analyze the text data from restaurant reviews using machine learning and deep learning capabilities, retaining the context of the data and attaining higher accuracies for better classification.
The first phase of the Study shares the idea of the business problem and highlights the research questions related to the problem. The next step showcases the importance of past work done, analyzes the models and techniques used to analyze the restaurant data, and sets a roadmap for the project.
Text data is highly unstructured data that needs to be dealt with using various text processing techniques to remove noise and help the machine learning models understand it. TF-IDF, count Vectorizer, and Word2Vec are used for the same, followed by the tokenization process.
The transformation stage and data preparation stage are the main stages on which the accuracies of these models rely, and hence, adequate analysis and vectorization techniques are required for the analysis to generate different architectures. The proposed solution showcases the capabilities of deep learning, machine learning, and hybrid architectures to identify and classify the reviews, generating labels that customers can use to analyze past reviews and make a decision. Also, the model can be used by restaurants to maintain quality and generate better services based on feedback management.

2 Related Work

Much research has been done in the field of finding information from text data. The Study focuses on the past work done to set the roadmap for the research that can guide and generate better results, focusing on various text mining capabilities and AI-based architectures.

2.1 Text Mining

The food domain uses text mining a lot to generate insights from the text data to facilitate applications like offering recipes, reinforcing food safety regulations, etc. The Study by (Xiong et al., 2024) highlights the use of text mining capabilities in understanding the food domain and showcasing its usages. Text classification is the most important domain of analyzing text

data, and the following is important in the food industry to showcase the association between food decisions and the health state of the food. Opinion mining is another application in the food industry used by fast-food restaurants, tea shops, and various food products to get consumer feedback and reviews and pinpoint various aspects like taste, texture, pricing, etc.(Ariyasriwatana & Quiroga, 2016). The same technologies can also be used for information extraction and NER (Named-Entity Recognition) to find the food components as well as the brand names, restaurant names, food menus, etc. Hence, text mining in the food industry will not only generate valuable information about food safety regulations but also provide valuable market intelligence. Also, the industry faces the issue of handling such a large and diverse dataset that showcases challenges of representativeness, completeness, privacy, etc.

(Gan et al., 2017) highlighted the use of text mining approaches along with multidimensional data to get the sentiment analysis for restaurant reviews. The Study used the data collected from the Yelp dataset challenge, having 335022 customer reviews from different businesses. The main focus of the research was to extract the restaurant data, and all the reviews, which were more than 100 words in length, were included in the research.(Mudambi & Schuff, 2010). The Study used a semi-supervised approach to find the topics that are highly correlated with positive as well as negative reviews. A research group studied and found the important words from the 2000 reviews, and these reviews were used to find the polarity based on the AFINN sentiment lexicons. Now, the sentiment is calculated by finding the weighted sentiment score based on the past reviews and the 5 variables. Two different models were built in the research, showcasing the relationship between consumer sentiment and star ratings. The next model added 5 extra variables to understand the relationship with star ratings. The Study showcased that these multidimensional variables emphasize the sentiment of the reviews as well as the star rating.

(Wen et al., 2024) used text mining to understand customer sentiment and product competitiveness on the online reviews dataset. The study used a total of 119,190 reviews collected from the e-commerce platform and cleaned from basic discrepancies like duplicates and missing values. The Study incorporated TF-IDF metrics to identify the major dimensions in the dataset. The study used clustering and sentiment analysis to track changes over the reviews. The Study aimed to find the eight critical dimensions to showcase customer concern and categorize them into four different quadrants for targeting and improvement. The Study focused on temporal analysis, and the sentiment and context can change over time. Also, the study highlighted that customers tend to become negative over time if they are not satisfied with the company.

One can understand the importance of text mining in the field of analyzing the review data. Now, the research will focus on making it better by using various machine learning architectures and understanding the importance of identifying patterns.

2.2 Machine Learning Approaches

Online reviews are a great source of information, whether it is a product or service. Generally, they are provided by short sentences or some star ratings that do not have much information for better decision-making. The Study by (Shin et al., 2022) worked on analyzing Google Maps data to generate insights about the food, price, service and atmosphere using sentiment analysis. The Study used Selenium web crawler to extract the data from the Internet and collected a total of 5427 reviews. The given corpus of reviews is then used to extract all the nouns with two or

more letters. The word taste showed up 1787 times in these reviews, followed by meat, waiter, etc. The given text data is then vectorized into a word list, and their counts are in the form of a dictionary. The given data is then vectorized using the TF-IDF vectorizer to get a better understanding, and the comments are labelled as 0 and 1. For ratings above 4 and 5, are 1 rest is 0. The data contained a total of 1257 negative reviews and 4170 positive reviews. The data is highly imbalanced and balanced via random sampling. The Study's main focus is to build up a dictionary of the most used words. Also, the Study suggests the use of machine learning libraries to give better results and focus more on the positive and negative aspects. The Study used a random forest algorithm, but much of the focus was not on accuracy as there are some positive reviews with negative words in them and vice versa. Hence, the Study highlighted the use of big data analytics and various deep learning models to get better results.

As can be seen, online reviews affect the restaurant business a lot. (Rita et al., 2023) In their paper, they focused on reviewing Michelin-starred restaurants. Social media is one of the most powerful tools of today's generation. People often use it to share their experiences about any product or service and also consider other reviews when making their purchase decisions. Michelin Star is one of the prestigious awards given to restaurants for the excellence of the restaurant within the city. This Study shows how people's sentiments towards four key aspects (food, service, ambience and price) change after a restaurant's award with a Michelin star. At the same time, we analysed 8,871 online reviews on Tripadvisor from 87 European restaurants extracted by a web crawler developed with Beautiful Soup. It is processed using Semantria to detect the sentiment in short words. Using the sentiment analysis tool, the Study finds that the overall sentiment in online reviews decreased after the restaurant got a Michelin star. Services were most negatively followed by food and ambience, but price sentiment shows significant increases. The study findings underscore the role of Michelin stars in shaping customer expectations. These stars, when awarded, significantly raise the bar for a restaurant's performance. If a restaurant fails to meet these heightened expectations, it is likely to result in more negative reviews. To maintain positive sentiments and reviews, Michelin-starred restaurants must continuously improve their food and service. They should also carefully manage high-end prices to align with perceived value.
Nature Language Processing (NLP) is a branch of Al that reads and understands human-generated text for sentiment analysis. This Study shows that sentiment analysis can reveal cultural differences in review content and customer perception. It also reveals that while the overall sentiment decreased after receiving a Michelin star, price sentiment for 1—and 3-star restaurants increased. Ambience sentiment is also decreased notably for 2-star restaurants. This finding suggests that Michelin star increase the reputation of the restaurant but also increases the customer expectation, which slightly increases the more negative sentiments towards those restaurants.

Mining customer opinions about restaurants have gained popularity in recent days because it is seen as impacting business growth and sustainability. The Study by (Sazzed, 2021) shows that there is only limited current sentiment analysis, which also often uses datasets of limited geographical regions such as the US, UK and China. Sentiment analysis or opinion mining refers to the process of identifying opinions or sentiments expressed (e.g., positive or negative) in a text document. Because user attitudes and preferences can be affected by many sociocultural factors, it is very important to have an annotated dataset from a diverse demographic. This research was conducted using a Bangla Resturants database with over 2300 reviews from all Bangladeshi restaurants.
This method uses a hybrid approach combining two methods, lexicon-based and machine learning (ML), to better analyse these text reviews. It also explores how demographic factors

affect linguistic features by comparing BanglaResturant with Yelp online reviews, which shows that demography significantly affects review language. The BanglaResturant dataset consists of reviews from Facebook pages of Bangladeshi restaurants. It contains a total of 2315 reviews, with 1702 positive and 613 negatives, in which positive reviews are recommended, and negative reviews are non-recommended. The research used four lexicon-based methods, which are VADER, Text Blob, LRSentiA, and SentiStrength and also five ML classifiers: Logistic Regression, Ridge Regression, SVM, Random Forest, and Extra Tree Classifier.
This Study categorizes reviews into Minimal Opinion Group (MOG), Fair Opinion Group (FOG), and Strong Opinion Group (SOG), with ML methods classifying MOG and FOG reviews. By comparing BanglaResturant with Yelp reviews it uncovers demographic influences, with Bangla Restaurant reviews being shorter and simpler. The research underscores the importance of considering demographic factors in sentiment analysis and proposes an inspiring direction for future studies: expanding this approach to other demographics and languages. This could potentially revolutionize sentiment analysis research by providing a more comprehensive understanding of user attitudes and preferences.

2.3 Deep Learning Approaches

After the machine learning approaches, the deep learning architecture shows a great relationship in finding the context of the data and is very helpful in restaurant review sentiment analysis. Sentiment analysis is one of the most widely applied topics of NLP. With big data analytics, the following has gained much momentum: it requires deep architectures to extract the information.(Khine & Aung, 2019)proposed the use of LSTM networks for the aspect-based sentiment analysis for the restaurant domain, focusing on aspects like price, reviews, etc, with respect to the restaurant domain. The Study implemented LSTM, which solves the problem of vanishing gradient descent for RNN models, and the attention mechanism gives attention to the ABSA task. The Study used data to extract the aspect it belongs to, and then the polarity of the review is found regarding the target. The Study incorporated MA-LSTM and the Sentic-Net model; both add attention mechanisms to LSTM, and the Sentic-Net is a knowledge base with 50,00 concepts. The study uses data from TripAdvisor with 20,000 sentences and tests with various other LSTM models, where the proposed model attained the maximum accuracy and outperformed these models by 3.2.

The user-generated content is increasing in the era of Web 2.0. The following environment has provided customers with a way to engage in social media activities, create content, and share their thoughts. (Li et al., 2021) In their paper, they have proposed the use of a Bi-GRU network, which is better than the LSTM networks and is used to increase sentiment classification, taking advantage of the attention mechanism based on the semantic dependency of the reviews. The study proposed a framework for using a web crawler to collect data from online platforms. The data is preprocessed, and all the null values are removed from the dataset. The given unstructured data is then converted into a structured format with the help of the Word2Vec model. The model uses CBOW and Skip-gram to learn the word representation and reduces the complexity of the algorithm. Also, the following is used to predict the context of the word. The proposed model of Bi-GRU uses the hidden layer to capture the formation of historical and future contexts. The same weight is multiplied by the input and the hidden layer at the previous point in time to ensure the context of the data. The model also introduced an attention mechanism, and the randomly crawled data of 35248 reviews and 130 stores, the model

attained a Recall of 93.77% and an F1-score of 89.45%, better than various baseline and deep learning models.

3 Research Methodology

3.1 KDD

KDD, also known as Knowledge Discovery in Databases, is a systematic process that is used to extract patterns from large datasets. As databases are growing in size, traditional approaches to extracting information from the data cannot be used. KDD helps extract patterns from large databases. The given process is a distributed process encompassing various stages. There are various classes where data mining can be used, such as Predictive Modelling, where one can solve the problem of Regression or classification analysis. Clustering talks about grouping the data based on similarity between the data points. KDD is also used in data summarization, dependency modelling and Time series analysis.(Fayyad & Stolorz, 1997). The framework of the KDD model is as follows:

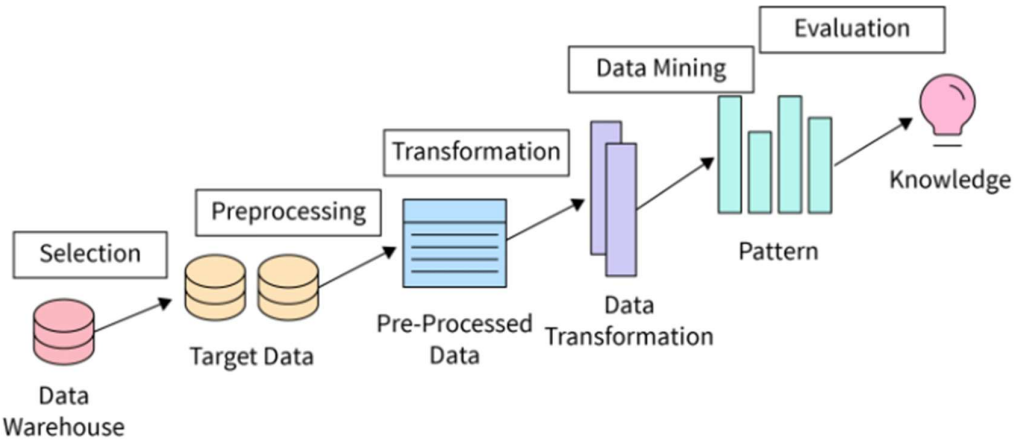

Figure 1 KDD Process

The following is a multi-step process that involves six stages to extract useful knowledge from databases, especially big data.

- **Data Selection –** The first phase of any data mining process is selecting the dataset from the larger databases. The following step involves getting information from the data warehouse or data streams to get the required data for analysis.
- **Data Preprocessing –** Once the dataset is available, the following gets converted into a cleaned format where all the noise, missing values and duplicates are removed from the data, making it clear and consistent. The following step is the most important in data mining.
- **Data Transformation –** Once the data is processed, the following is transformed in a format that the model can analyze. Here, the data is converted into another format. Normalization, vectorization,n and aggregation are some of the steps that are followed here.

- **Data Mining –** Now, the prepared dataset is used for model building, where the models are applied to the given dataset to find the hidden patterns from the database. Based on the type of problem, like clustering, classification, etc, the models are built.
- **Evaluation –** Once the models are built, they are evaluated based on various metrics to evaluate the patterns and the significance of the model.
- **Knowledge Presentation –** The last step is to use this knowledge and present it to humans for better decision-making. Combining knowledge with human intuition can generate great results.

As seen, the given process is standard, and the following needs to be modified to be used on the text data. As the text data is highly unstructured, the following cannot be used there.

3.2 KDT

Knowledge discovery in text databases is a modified version of KDD that specializes in extracting patterns from unstructured text data. The following was built to satisfy the need for vast amounts of text data generated across various platforms, including social media, blogs, publications, etc. Text categorization is the simplest and most acceptable procedure that can be used for the implementation of KDT, and the research focuses on classifying the sentiment behind our view using the same methodology.(Feldman & Dagan, 1995). The following methods include various disciplines like data mining, NLP (Natural language Processing), Information Extraction, Sentiment Analysis and Topic Modelling.

- **Natural Language Processing –** Natural language processing is a field of AI that is focused on enabling computers to understand, interpret and generate the human language. The following combines techniques from computer science, linguistics and machine learning to analyze the text data. The NLP deals with getting the grammar structure of the text, the meaning of words, and analyzing and generating text data.(Shankar & Parsana, 2022)
- **Information Extraction –** Information extraction represents getting structured information from unstructured data like text documents, web pages and larger databases. It deals with finding the entities and relationships and presenting them in a structured format. One of the most common examples is entity recognition, which allows one to classify entities such as persons, organizations, locations, etc.(Small & Medsker, 2014)
- **Sentiment Analysis –** Sentiment analysis is the subfield of NLP that deals with extracting information from the text and determining the sentiment behind the text, using the sentiment behind it and labelling it positive, negative, and neutral. It can be used to find the polarity of the data and is also used for emotion detection.(Saxena et al., 2022)
- **Topic Modelling –** Topic modelling is also a technique used in text mining to find themes from the data. The following technique is used to infer the main topics from the text and to organize and understand the large volume of text databases. The most common implementation of topic modelling is LDA, which considers each document as a mixture of topics and each topic as a distribution of words. (Churchill & Singh, 2022)

3.3 Design Specification

The given design specification for implementing the research is highlighted below. That shares the flow of information to extract the sentiment behind it and the techniques used in the research to clean the text data.

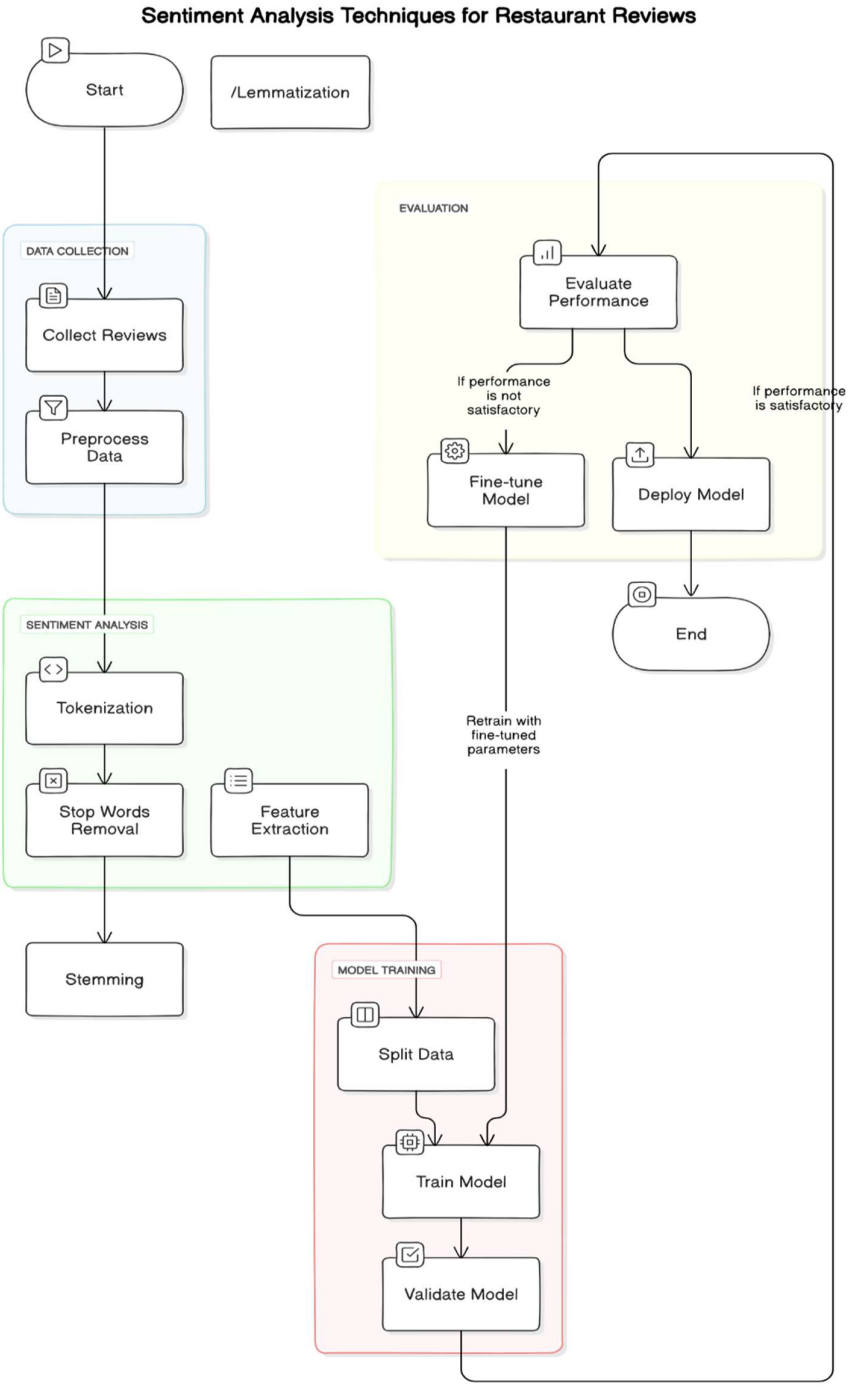

Figure 2 Flowchart

4 Implementation

4.1 Introduction

The chapter demonstrates how the KDT is implemented on the text data to find the sentiment analysis based on reviews from restaurant data and what different text processing steps need to be taken to build various machine learning and deep learning architectures.

4.2 Data Selection

The dataset used in the research is taken from the Zenodo website, which has data available from various restaurants in Dublin across 65 locations.(Basheer, 2019) the data dictionary of the given dataset is as follows:

Table 1 Data Dictionary

Column Name	Description
Restaurant ID	All those IDs correspond with different restaurants that are identified by the unique ID for ease of reference. The count of restaurants is 211 in total.
Location ID	The IDs of the respective locations are provided for every restaurant. In total, this data contains information regarding 65 such places.
Cuisine	Shows the type of cuisine in each restaurant.
Price Range	The average price for dining in the restaurant. It has 3 categories: €30 and under | €31 to €50 | €51 and over
Food Rating	The rating given by users between 1 to 5 for the quality of the food in the restaurant.
Ambience Rating	The rating given by users between 1 to 5 for the ambience in the restaurant.
Overall Rating	The rating given by users between 1 to 5 for the restaurant.
Restaurant Rating	The average rating of each restaurant is based on general opinion.

Review	This field shows the the user reviews for each restaurant. Reviews have been cleaned to remove punctuations, emojis and special characters.
Service Rating	It shows the given a rating by customers between 1 and 5 for the quality of service in the restaurant.
Review Sentiment	The overall sentiment of the user review, i.e. either Positive or Negative.

The dataset contains a total of 10,000 records and 12 columns, as seen in the data dictionary. Data analysis on the given dataset is described as follows:

- Sentiment Score density is high where the food rating is high, and the sentiment score is low for the food ratings having low density. Some negative reviews reach a food rating of 5, and the rest are below that only.

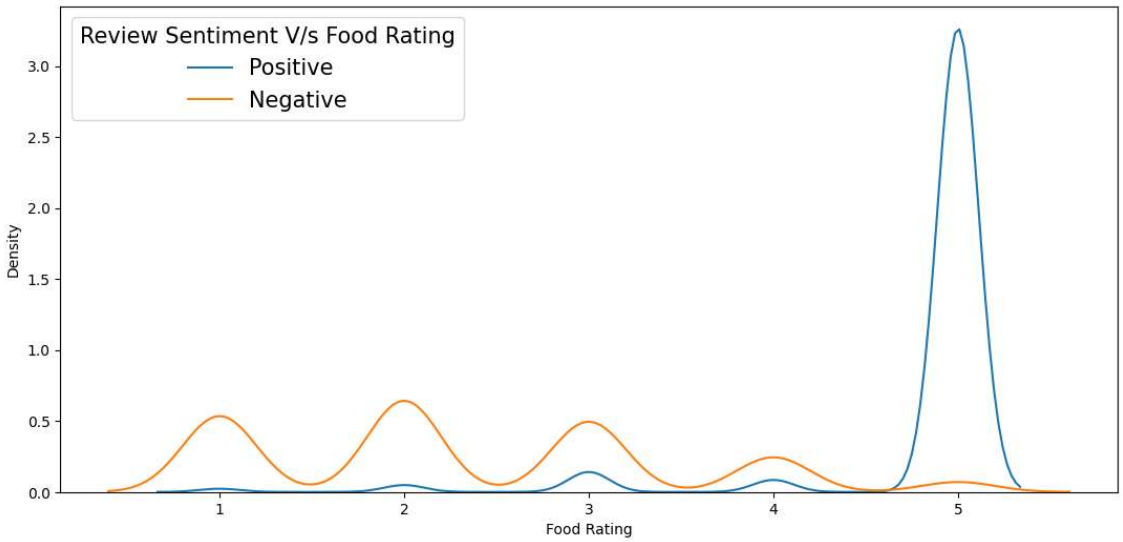

Figure 3 Sentiment V/s Food Ratings

- The same patterns are observed for the Service Rating and Restaurant Rating, where the density of ratings is high for 5 stars for services. However, there are restaurants with negative restaurant ratings that still have sentiment ratings of 5. Hence, we are highlighting a mix-match of the sentiment w.r.t to restaurant ratings.
- From Figure 6, one can see that the most popular cuisines are Irish, followed by Italian and European, and the least is Cocktail Bar and Pubs.

These are the basic data analyses of the data that one can understand. Now, the study will shift towards analyzing the text data and focus on its cleaning.

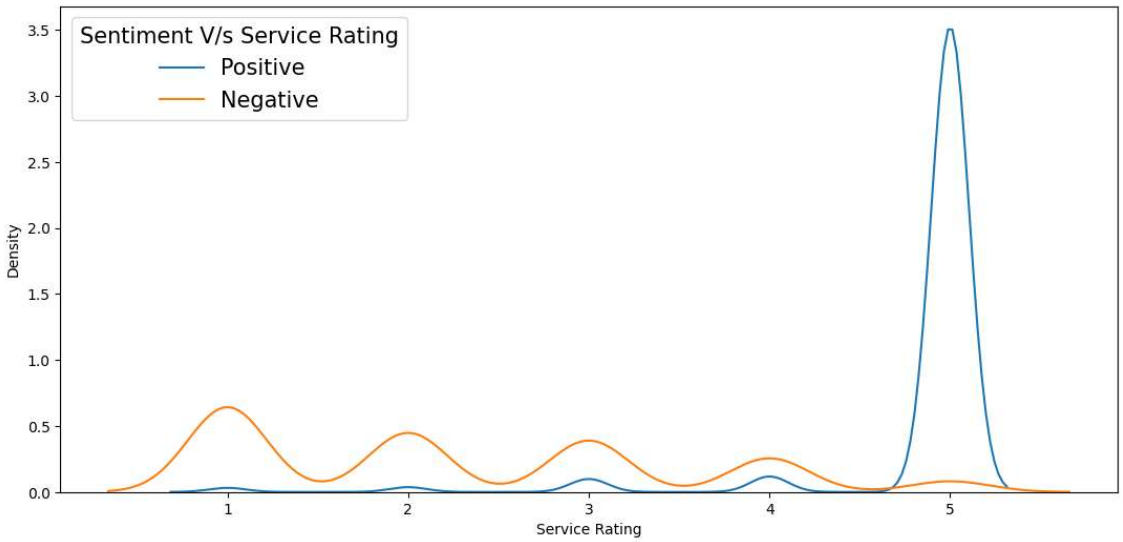

Figure 4 Sentiment V/s Service Rating

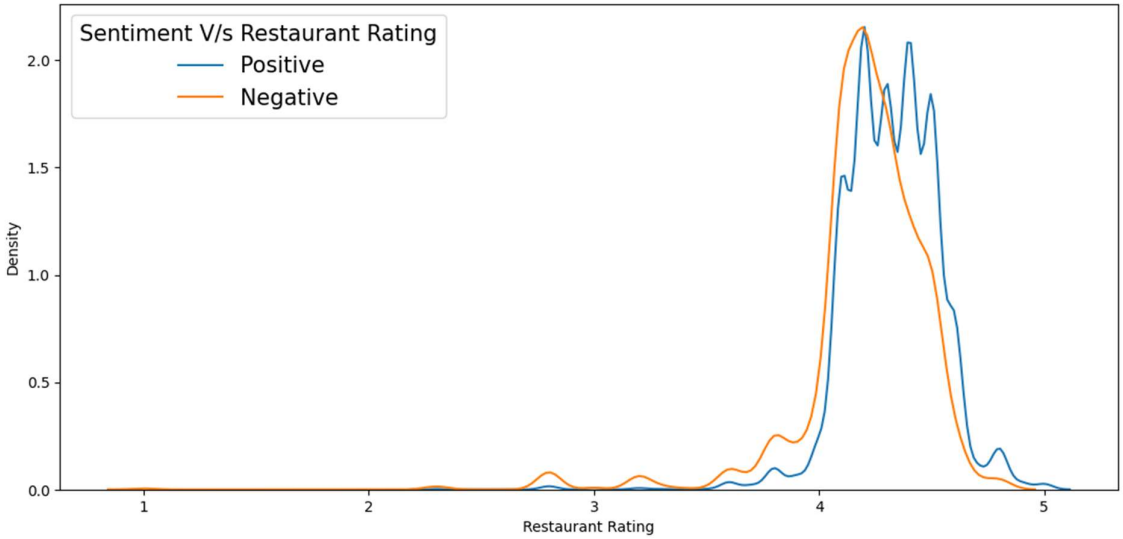

Figure 5 Sentiment V/s Restaurant Rating

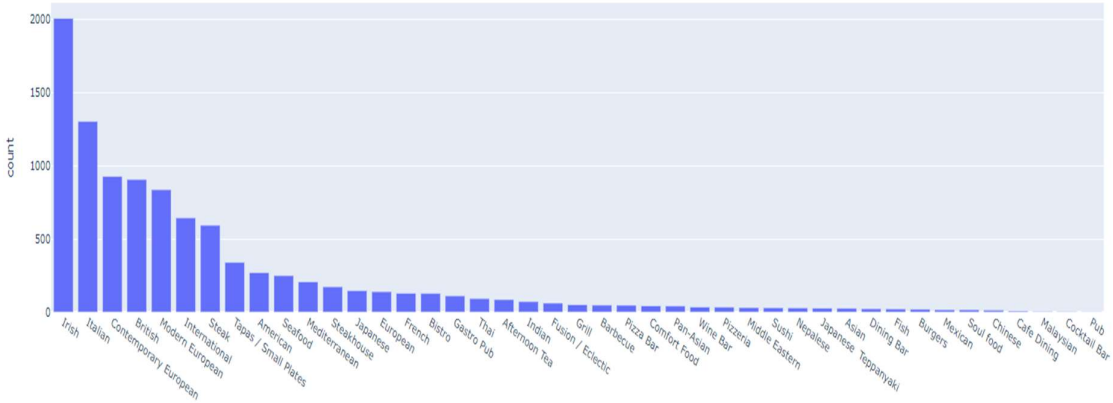

Figure 6 Count of Cuisines

4.3 Natural Langauge Understanding

The next step is to implement the KDT and focus on the text data or categorize the text data based on the reviews. The following part of analyzing the text data comes under the NLU or Natural Language Understanding. The following step makes understanding the input variable easy and helps analyze the language. It also helps in finding the sentiment of the sentence. The steps involved in data cleaning under NLU are discussed as follows:

4.3.1 Text Standardization

The first step in the text analysis is to convert the data into lower cases because Python is a case-sensitive language, and the following conversion makes the interpretation easier.

4.3.2 Punctuation Removal

Any text data used on social media apps or in reviews may or may not contain some special characters and emojis. These characters do not make any sense in the text data and need to be removed. The research utilized regular expressions to remove the extra words except alphabets from the data.

4.3.3 Stopwords Removal

Once the data is cleaned and standardized, the next step is to remove the stopwords from it. Stopwords are meaningless words from a language that help build the meaning of the sentence and do not have any sentiment. Removing the stopwords helps the language models improve the accuracy and relevance of the language as per context.

4.3.4 Stemming and Lemmatization

The most important step in the NLU is stemming and lemmatization of the text data. The given preprocessing step helps maintain the context for grammatical reasons. Both of these techniques work for the same purpose of reducing the word to the base form.(What Are Stemming and Lemmatization? | IBM, 2023)

- **Stemming:** Stemming process removes the prefixes and suffixes from the word to obtain the root word or stem of the word. The following process works based on some rules and does not consider any grammatical context.
- **Lemmatization:** Lemmatization, in contrast, is an advanced procedure that considers the word's context and meaning. Reconstructing inflected forms into their base or dictionary forms, called lemmas, calls for an extensive vocabulary as well as morphological study.

Once the data is cleaned, certain word clouds are drawn for both the Positive and Negative Sentiment, finding the top 50 words. Figure 7 showcases the top 50 words with Positive Sentiment, and Figure 8 showcases the words with negative sentiment. One can observe keywords like recommend, highly, staff, food, etc., as the highlighted keywords in positive sentiment and the same way for negative sentiment. The keywords are staff, time, waiter, food etc.

Figure 7 Positive Words

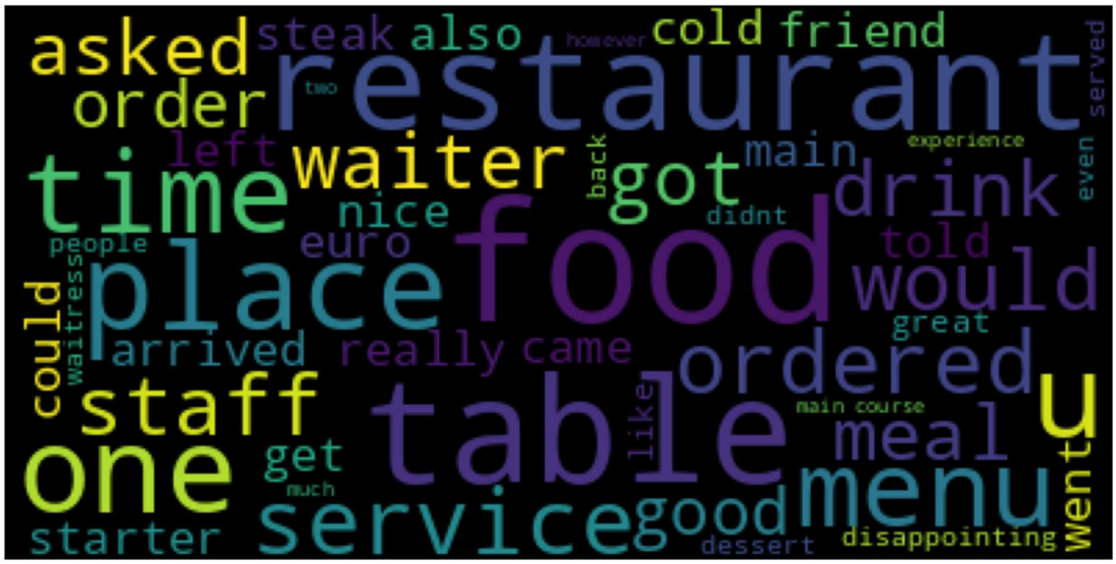

Figure 8 Negative Words

4.3.5 Vectorization

Three different vectorization techniques are used in the research to convert the given text data into numerical format. The following transformation helps in analyzing the text data effectively.

- **Count Vectorizer:** Count Vectorizer is one of the simplest methods that is used to convert the collection of text documents into a matrix containing the token counts. The following counts the number of times a word occurs and showcases the presence of a word without its significance. The method counts the number of times a word comes up in a document, making every word a feature in the corpus. The output of the given

method is a matrix containing all the words as columns and counted as rows. For a large dataset, the following becomes a very large matrix.

- **TF-IDF:** Term Frequency, Inverse Document Frequency is again one of the renowned matrices used in the NLP. The following matrix showcases the importance of a word in the document as compared to multiple documents. The following matrix is calculated using two factors. Here, the Term Frequency measures how frequently a term occurs in the document and is calculated using the number of times a term appears in the document divided by the total number of terms in the document. The IDF term measures the importance of a term by calculating the logarithm of the total number of documents divided by the number of containing a particular term. The following matrix helps in reducing the weight of most common words and provides significance to the words used least. (Apply the (TF-IDF Vectorization Approach, 2022; Kilmen & Bulut, 2022)
- **Word2Vec:** The last matric used in the research is Word2Vec. The following are advanced techniques that use the deep learning architecture neural net to generate the word embeddings. The output is a dense vector representation of words where similar words are placed close to one another in vector space. CBOW, also known as Bag-of-words, is one architecture that predicts the target word based on its context and another architecture that predicts the surrounding word based on the target word, and it is Skip-Gram. The following method is superior in finding the word similarity or sentiment analysis as it retains the contextual and semantic information of the word.

4.4 Modelling

Once the data is prepared, the next step is to build the model, and all the possible combinations are used to build the model with the output of given vectorization techniques, building a total of 15 combinations. The models used in the research are:

4.4.1 Logistic Regression

Logistic Regression is one of the base models used in the classification task. The model is famous for binary classification. The given model is inspired by the statistical sigmoid function that builds the relationship between the dependent and dependent variables by estimating the probability. The output of logistic Regression is always between 0 and 1, and it tells the likelihood of a class label. The following model is used as the baseline model in NLP. The given model is used in the research because, with this model, one can have the text metrics represented in the form of weights. The model is simple to use and highly efficient compared to other models.(Natural Language Processing (NLP) for Sentiment Analysis with Logistic Regression, n.d.)

4.4.2 Naïve Bayes

The Naïve Bayes classifier is a probabilistic classifier that relies on the Bayes theorem. This is a probabilistic classifier that assumes all the features are independent of one another. The algorithm is an essential part of generative learning and does not acquire an understanding of the significance of individual components during the training of the model. A model is utilized in the text classification. Harry R Felson and Robert M. Maxwell were pioneers in utilizing Naïve Bayes with text vectors for text classification models. Bayes theorem is also applied in the analysis of text data.(How to Use Naive Bayes for Text Classification in Python?, n.d.)

4.4.3 Xg-Boost

Xgboost model is known for its high performance and is a distributed, open-source machine learning library that uses the advantage of gradient-boosted decision trees and the advantage of system software and hardware. The model is known for its efficiency and ability to scale well on large datasets. The base of the algorithm is gradient-boosted decision trees, which are a boosting algorithm that uses the approach of gradient descent to minimize loss. The gradient boosting algorithm uses parallel and distributed computing for the model building and provides built-in regularization to get better results. As the given vectorized data is highly sparse, the xgboost model can show great results.

4.4.4 Random Forest

Random Forest is one of the powerful machine-learning algorithms that is backed by the concept of bagging. Developed by Leo Breiman and Adele Cutler, it is one of the popular algorithms in supervised machine learning and is well-suited for regression and classification-based problems. The given popularity applies in the fields of text analytics, sentiment analysis, etc. The model can handle large and sparse datasets very easily and is robust to any noise in the dataset. The model provides estimates of feature importance, helping in finding the most influential words in finding the sentiment. As the model is an ensemble model, it controls the overfitting of the model on the dataset and is a strong choice for usage.(What Is Random Forest? | IBM, n.d.)

4.4.5 LSTM

Long Short-Term Memory, or LSTM, is a special type of neural network grown from the RNN model. Responsible for learning the long-term dependencies. The model was introduced by (Hochreiter Schmidhuber, 1997). LSTM models solve the problem of long-term dependency, where the layers especially interact with one another. The model works really well with sequential data like time series and text data. The major components of the LSTM model are:

- Cell State that carries information across different time intervals
- Forget gate to forget the information from a cell state and
- Output gate that decides the next hidden state.

All these gates control the flow of information and help the model retain, forget, and update information to make long-term dependencies.(Understanding LSTM Networks -- Colah's Blog, n.d.)

5 Evaluation

The researcher aims to build a classifier that can classify the sentiment based on the text data and extract the maximum information available from the text data for the given task. The study has focused on building various text metrics and models to try out different combinations to get the best benchmark model. Different evaluation metrics have been used in the research to find the best model as per KDT. To evaluate the model and justify the model's ability to understand all the objectives. The metrics used are:

- **Accuracy:** Accuracy is a classification metric that tells the proportion of a number of correctly predicted sentiments out of the total number of predictions made. It is a base metric used in sentiment analysis and validates the model in the best possible manner when the data is balanced.

- **ROC-AUC:** Receiver Optimization Characteristics, Area under the curve is the second metric used in the research that focuses on providing the true picture of the model as the following metric tells how good the model is in identifying a sentiment Positive and Negative. The metric is calculated by sensitivity and 1-specificity, where sensitivity means the proportion of actual positive values that are correctly identified by the model, and specificity, which tells the proportion of actual negative sentiment correctly identified by the model. The better the AUC score is, the better the model is in classification.
- **Classification Report:** The classification Report is the sum up of the above two metrics and provides a detailed analysis of the performance of the classification model. The report includes metrics like precision (how good the model is in identifying true positive out of all the True Positives and False Positives), Recall (Sensitivity), F1-score (Harmonic mean of Precision and Recall), and support (number of occurrences of each class in the dataset)
- **Confusion Matrix:** Confusion Matrix details the correct and incorrect classification by providing details like True Positives, True Negatives, False positives, and False Negatives. The matrix provides all these details in the form of a 2*2 matrix in case of binary classification to get a clear picture of the results.

Out of all the metrics available, the major metrics used are accuracy scores and AUC scores, as they provide a clear picture. A confusion matrix is used just as a way to visualize the misclassifications that are happening. All the models performed really well with Count Vectorizer and TF-IDF, and the LSTM model attained the maximum accuracy of 0.959 with Count-Vectorizer for the given problem, and the same model attained the least accuracy of 0.498 with TF-IDF however, the logistic Regression is the second model of choice for the problem with TF-IDF with an accuracy of 0.944

Table 2 Accuracy Scores

Model	Vector Matrix	Accuracy
Logistic Regression	Count-Vectorizer	0.947
	TF-IDF	0.944
	Word2Vec	0.522
Naïve Bayes	Count-Vectorizer	0.937
	TF-IDF	0.939
	Word2Vec	0.523
Xg-Boost	Count-Vectorizer	0.938
	TF-IDF	0.936
	Word2Vec	0.522
Random Forest	Count-Vectorizer	0.937
	TF-IDF	0.921
	Word2Vec	0.522
LSTM	Count-Vectorizer	0.954
	TF-IDF	0.498
	Word2Vec	0.505

Table 3 AUC Scores

Model	Vector Matrix	AUC
Logistic Regression	Count-Vectorizer	0.947
	TF-IDF	0.944
	Word2Vec	0.519
Naïve Bayes	Count-Vectorizer	0.937
	TF-IDF	0.939
	Word2Vec	0.521
Xg-Boost	Count-Vectorizer	0.938
	TF-IDF	0.936
	Word2Vec	0.519
Random Forest	Count-Vectorizer	0.937
	TF-IDF	0.921
	Word2Vec	0.519
LSTM	Count-Vectorizer	0.989
	TF-IDF	0.50
	Word2Vec	0.515

The same configuration achieves the maximum AUC score of 0.989 for the given model, making it the best model for sentiment analysis where the given model can relate to the context of the data and classify the sentiment. The model has also outperformed in various text classification tasks in research and is versatile on text datasets. Laos, the model can be combined with various other models to enhance the performance further. Combining CNN for feature extraction results in higher accuracies. (Gondhi et al., 2022)

6 Conclusion and Future Work

The research presented a standard data mining approach, KDT (Knowledge Discovery in Text), to classify whether the review was positive or not by applying natural language processing to the restaurant review dataset. A total of 3 Text metrics are used to understand which metric the models can give the maximum accuracy and whether machine learning is suitable or deep learning models are best for learning the context of the data. The study followed the standard process of data cleaning using NLU and removed all the noise from the data.

The central theme of the research is to understand the text data and find the best model for various permutations of models where the LSTM model attained the maximum accuracy and ROC score with Count-Vectorizer followed by logistic Regression with TF-IDF. However, surprisingly low results were obtained with the Word2Vec model. The research tries to answer all the research questions. **RQ1** The given models with different vectorization techniques can classify the sentiment in a better manner, and examining the data provides a better insight into the data. **RQ2,** the given models are hard to explain as they contain a sparse matrix. However, techniques like Layerw-se Propagation relation (LRP) and the attention mechanism of BERT can be used to explain it in future advancements, helping customers and restaurants make better decisions. However, the research highlights the most common keywords to answer some questions.**RQ3** The model accuracy changes over various text metrics in some cases. The Count Vectorizer gave the best results, and TF-IDF also gave the best results comparatively, and there is a huge gap with the Word2Vec model.

The study provides new knowledge in the given field by building a robust sentiment analysis framework that can help understand customer preference based on sentiment. The design of the study is appropriate for these research questions, and it dwells on exploring data mining capabilities to extract patterns from text data using KDT based on aspect-based sentiment analysis. The data collected from Zenodo is an open-source dataset and provides a rich source of information for the research objectives. Also, the data is a public review dataset and is accessible to everyone, so there is no privacy concern with the model. As the review language may add bias in the model based on region, the approach of aspect-based analysis removes that, and there are not many ethical concerns.

The model attained the best accuracy, but for further improvements in the future, the study will focus on testing various deep learning classifiers like more advanced models BERT. Also, the study will try to capture a diversity of information by collecting datasets from review websites like Yelp or Google Reviews. A sentiment lexicon can be built for the given domain and can be used to detect sentiment more accurately. Advanced techniques like Word2vec and Glove can be used for better word embeddings, and models can provide real-time analysis along with explainability for better decision-making. Also, user feedback can be incorporated into the training process to refine the model with the best accuracy.

7 References

Apply the TF-IDF Vectorization Approach. (2022). https://openclassrooms.com/en/courses/6532301-introduction-to-natural-language-processing/8081363-apply-the-tf-idf-vectorization-approach

Ariyasriwatana, W., & Quiroga, L. M. (2016). A thousand ways to say' Delicious!'—Categorizing expressions of deliciousness from restaurant reviews on the social network site Yelp. *Appetite*, *104*, 18–32. https://doi.org/10.1016/J.APPET.2016.01.002

Basheer, A. K. A. (2019). *Social website reviews and ratings of Dublin restaurants situated across 65 locations*. https://doi.org/10.5281/ZENODO.3356793

Churchill, R., & Singh, L. (2022). The Evolution of Topic Modeling. *ACM Computing Surveys*, *54*(10s), 1–35. https://doi.org/10.1145/3507900

Fayyad, U., & Stolorz, P. (1997). Data mining and KDD: Promise and challenges. *Future Generation Computer Systems*, *13*(2–3), 99–115. https://doi.org/10.1016/S0167-739X(97)00015-0

Feldman, R., & Dagan, I. (1995). *Knowledge Discovery in Textual Databases (KDT)*. www.aaai.org

Gan, Q., Ferns, B. H., Yu, Y., & Jin, L. (2017). A Text Mining and Multidimensional Sentiment Analysis of Online Restaurant Reviews. *Journal of Quality Assurance in Hospitality & Tourism*, *18*(4), 465–492. https://doi.org/10.1080/1528008X.2016.1250243

Gondhi, N. K., Chaahat, Sharma, E., Alharbi, A. H., Verma, R., & Shah, M. A. (2022). Efficient Long Short-Term Memory-Based Sentiment Analysis of E-Commerce Reviews. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/3464524

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/NECO.1997.9.8.1735

How to Use Naive Bayes for Text Classification in Python? (n.d.). Retrieved August 8, 2024, from https://www.turing.com/kb/document-classification-using-naive-bayes

Khine, W. L. K., & Aung, N. T. T. (2019). Applying Deep Learning Approach to Targeted Aspect-based Sentiment Analysis for Restaurant Domain. *2019 International Conference on Advanced Information Technologies (ICAIT)*, 206–211. https://doi.org/10.1109/AITC.2019.8920880

Kilmen, S., & Bulut, O. (2022). *Text Vectorization Using Python: TF-IDF*. https://okan.cloud/posts/2022-01-16-text-vectorization-using-python-tf-idf/

Li, L., Yang, L., & Zeng, Y. (2021). Improving sentiment classification of restaurant reviews with attention-based bi-gru neural network. *Symmetry*, *13*(8). https://doi.org/10.3390/sym13081517

Local Consumer Review Survey 2024. (n.d.). Retrieved August 2, 2024, from https://www.brightlocal.com/research/local-consumer-review-survey/#

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on amazon.com. *MIS Quarterly: Management Information Systems*, *34*(1), 185–200. https://doi.org/10.2307/20721420

Natural Language Processing (NLP) for Sentiment Analysis with Logistic Regression. (n.d.). Retrieved August 8, 2024, from https://blog.mlq.ai/nlp-sentiment-analysis-logistic-regression/

Rita, P., Vong, C., Pinheiro, F., & Mimoso, J. (2023). A sentiment analysis of Michelin-starred restaurants. *European Journal of Management and Business Economics*, *32*(3), 276–295. https://doi.org/10.1108/EJMBE-11-2021-0295

Saxena, A., Reddy, H., & Saxena, P. (2022). *Introduction to Sentiment Analysis Covering Basics, Tools, Evaluation Metrics, Challenges, and Applications* (pp. 249–277). https://doi.org/10.1007/978-981-16-3398-0_12

Sazzed, S. (2021). A Hybrid Approach of Opinion Mining and Comparative Linguistic Analysis of Restaurant Reviews. *International Conference Recent Advances in Natural Language Processing, RANLP*, 1281–1288. https://doi.org/10.26615/978-954-452-072-4_144

Shankar, V., & Parsana, S. (2022). An overview and empirical comparison of natural language processing (NLP) models and an introduction to and empirical application of autoencoder models in marketing. *Journal of the Academy of Marketing Science*, *50*(6), 1324–1350. https://doi.org/10.1007/s11747-022-00840-3

Shin, B., Ryu, S., Kim, Y., & Kim, D. (2022). Analysis on Review Data of Restaurants in Google Maps through Text Mining: Focusing on Sentiment Analysis. *Journal of Multimedia Information System*, *9*(1), 61–68. https://doi.org/10.33851/JMIS.2022.9.1.61

Small, S. G., & Medsker, L. (2014). Review of information extraction technologies and applications. *Neural Computing and Applications*, *25*(3–4), 533–548. https://doi.org/10.1007/s00521-013-1516-6

Sukanya, M., & Biruntha, S. (2012). Techniques on text mining. *Proceedings of 2012 IEEE International Conference on Advanced Communication Control and Computing Technologies, ICACCCT 2012*, 269–271. https://doi.org/10.1109/ICACCCT.2012.6320784

Understanding LSTM Networks -- colah's blog. (n.d.). Retrieved August 8, 2024, from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Wen, Z., Chen, Y., Liu, H., & Liang, Z. (2024). Text Mining Based Approach for Customer Sentiment and Product Competitiveness Using Composite Online Review Data. *Journal of Theoretical and Applied Electronic Commerce Research*, *19*(3), 1776–1792. https://doi.org/10.3390/jtaer19030087

What Are Stemming and Lemmatization? | IBM. (2023). https://www.ibm.com/topics/stemming-lemmatization

What Is Random Forest? | IBM. (n.d.). Retrieved August 8, 2024, from https://www.ibm.com/topics/random-forest

Xiong, S., Tian, W., Si, H., Zhang, G., & Shi, L. (2024). A Survey of the Applications of Text Mining for the Food Domain. *Algorithms 2024, Vol. 17, Page 176*, *17*(5), 176. https://doi.org/10.3390/A17050176