



**Customer Experience: Journey Disruption Forecasting- Sentiment Analysis
of Customer Feedback Using Machine Learning and Deep Learning**

Prepared By

Name: Nouf Alotaibi

ID No:243001489

Submitted in fulfillment of the requirements for the degree Master of Science in
Artificial Intelligence

At

Collage of Informatics

Midocean University

UAE

Supervisor Name: Dr. Hagar Saleh

Collage of Informatics, Midocean University, Comoros

Email:243001489@st.midocean.edu.km

Acknowledgements

I would like to express my deepest appreciation to Dr. Hagar Saleh for her exceptional guidance, continuous support, and invaluable insights throughout the development of this thesis. Her expertise, constructive feedback, and unwavering encouragement have significantly shaped the direction and quality of this research. I am sincerely grateful for her time, dedication, and commitment to ensuring the academic rigor and clarity of this work.

My gratitude also extends to the faculty and staff of the Midocean University for providing a supportive academic environment that enabled the successful completion of this study. I further wish to acknowledge all colleagues, peers, and individuals who contributed directly or indirectly to the analysis, discussions, and feedback that enriched this thesis.

Finally, I would like to thank my family and everyone who believed in my ability to complete this work, offering motivation, patience, and encouragement throughout this journey. Their support has been instrumental in reaching this milestone.

Abstract

The growing adoption of customer experience analytics has increased the importance of sentiment analysis as a key mechanism for understanding customer feedback. The emergence of digital communication tools such as chat services, email, and other online platforms has driven a strong demand for an efficient method for sentiment classification of large volumes of customer complaint data.

This thesis proposes a model that combines the traditional and deep learning models using the transformer architecture. This provides a holistic approach to classifying customer complaint sentiment. The model was tested on the Customer_Complaints_Sentiment dataset on Kaggle, which consists of customer complaint text samples classified as either positive or negative. The natural language processing techniques include text processing, normalization, feature selection, and label preparation.

Several machine learning algorithms were applied, including Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost), using Word Count and TF-IDF features. Additionally, transformer networks, specifically BERT and DistilBERT, were applied to capture the context/semantic information present in the complaint descriptions. The performance metrics for all applied machine learning algorithms are Accuracy, Precision, Recall, F1-score, and a confusion matrix.

According to the experimental results, the model's performance largely depends on the feature representation. TF-IDF performed better than Word Count in the traditional machine learning approach. Among the traditional approaches, XGBoost achieves the best results (Accuracy = 0.7685, F1-score = 0.7379). Results for the transformer-based approaches are also impressive. BERT has the best performance (Accuracy = 0.7486, F1-score = 0.7322). Similarly, DistilBERT performed comparably to BERT but with less complexity.

Further analysis of the confusion matrix also shows that negative-sentiment detection with transformer models outperforms conventional models, which predominantly classify instances based on the majority class. In line with this, results affirm the efficacy of contextualized models in processing implicitly expressed consumer dissatisfaction in more complex contexts. This research contributes to the existing literature by presenting a comprehensive comparative study of a variety of machine learning and transformer models on a customer complaint dataset, which has been relatively underexplored compared to general consumer review datasets. The approach presented here has key implications for businesses seeking to leverage sentiment analysis to improve customer satisfaction.

Keywords: Sentiment Analysis, Customer Feedback Classification, Machine Learning, Deep Learning, Transformer Models.

Contents

Acknowledgements	2
Abstract	3
List of Abbreviations	5
List of Figures.....	6
List of Tables	7
Chapter 1 Introduction	8
Introduction	8
Problem Definition.....	8
Thesis Objectives	9
Thesis Contribution:.....	9
Thesis Organization:	10
Chapter 2 Overview	11
Chapter 3 Related Work	19
Chapter 4 Proposed Approach	21
A. Data Collection.....	21
B. Data Preprocessing	22
C. Feature Representation Techniques	23
D. Model Training:	23
E. Evaluation Metrics.....	24
Chapter 5 Experimental Results and Discussion	25
Chapter 6 Conclusion and Future Work.....	31
References	32

List of Abbreviations

- **Acc.:** Accuracy
- **Prec.:** Precision
- **Rec.:** Recall
- **F1:** F1-Score
- **ROC-AUC:** Receiver Operating Characteristic – Area Under the Curve
- **ML:** Machine Learning
- **DL:** Deep Learning
- **NLP:** Natural Language Processing
- **AI:** Artificial Intelligence
- **LR:** Logistic Regression
- **DT:** Decision Tree
- **RF:** Random Forest
- **KNN:** K-Nearest Neighbors
- **LCTM:** Latent Class Topic Model
- **GRU:** Gated Recurrent Unit
- **BERT:** Bidirectional Encoder Representations from Transformers
- **SVC:** Support Vector Classifier
- **TF-IDF:** Term Frequency–Inverse Document Frequency
- **BoW:** Bag of Words
- **EDA:** Exploratory Data Analysis
- **VOC:** Voice of Customer
- **TP:** True Positive
- **TN:** True Negative
- **FP:** False Positive
- **FN:** False Negative
- **CPU:** Central Processing Unit
- **GPU:** Graphics Processing Unit
- **API:** Application Programming Interface

List of Figures

Fig. 2.1 – Artificial Intelligence (AI)
Fig. 2.2 – Natural Language Processing (NLP)
Fig. 2.3 – Machine Learning workflow
Fig. 2.4 – Deep Learning Architecture
Fig. 2.5 – Transformer (BERT) Architecture
Fig. 2.6 – Text Preprocessing Techniques
Fig. 2. 7 Feature Engineering
Fig. 2.8 Feature Representation Techniques
Fig. 2.9 – Classification Techniques Workflow
Fig. 2.9.1 - Decision Tree (DT)
Fig. 2.9.1 - Random Forest (RF)
Fig. 2.9.1 - KNN
Fig. 2.9.1 – SVC
Fig. 2.10 - Feature Representation
Fig. 2.11 - Evaluation Metrics
Fig. 2.11 – Sentiment Analysis Conceptual Workflow
Fig. 4 – Methodology
Fig.5.2.1 Confusion Matrix- Word Count Models
Fig.5.3.1 Confusion Matrix- TF-IDF Models
Fig.5.4.1 Confusion Matrix-Transformer Models BERT
Fig.5.4.1 Confusion Matrix- Transformer Models DistilBERT

List of Tables

Table 4.A – Data Example

Table 4.B – Data Preparation

Table 4.E – Evaluation Metrics

Table 4.E – Confusion Metrics

Table 5.2 – ML Using Word Count Results

Table 5.3 – ML Using TF-IDF Results

Table 5.4 – Transformer Based Models Results

Chapter 1 Introduction

Introduction

One such task that has been significantly impacted by the rapid evolution of NLP, as a result of the burgeoning use of text in different forms, such as emails, live chat communications, surveys, and customer grievance portals, is sentiment analysis. Due to its increasing use in comprehending customer sentiment, as well as in determining service quality, the automated sentiment classification task has played a pivotal role in enabling customer experience analytics[1][2].

Although there has been rapid development in the field of machine learning (ML) as well as deep learning (DL), sentiment analysis of real customer grievances still proves a difficult task. Customer grievances always comprise informal language, writing styles, terminology, as well as noisy structure that makes the application of traditional machine learning classifiers such as Logistic Regression, Naïve Bayes, or Support Vector Machines less effective[3]. Due to this, more advanced models of deep learning such as LSTM, Bi-LSTM, or GRU, having capabilities of dealing with sequential dependencies, gained popularity[4].

Recently, the emergence of transformer-based models like BERT has shown improvements in sentiment classification tasks through self-attention capabilities [5]. However, there is a dearth of studies that focus on comparisons of traditional machine learning, deep learning, and transformer-based models with realistic organizational complaint datasets that embody linguistic noise.

This thesis endeavors to fill this gap by testing a variety of ML, as well as DL, models, together with BERT, on a public sentiment collection[6]. By incorporating a thorough NLP processing system, this thesis will offer perspective on model efficiency within a live customer service scenario.

Problem Definition

Organizations have long struggled with accurately and timely detection of customer sentiment. Often, they realize where the problem lies only when the discontent has been expressed in the form of grievances and service requests, and by then, the delay in receiving negative feedback has already occurred. Despite collecting a substantial amount of textual data, organizations typically generate sentiment insights only during scheduled reporting cycles rather than in real time[7][8].

Moreover, the data related to customer opinions and feedback are unstructured and characterized by diverse writing styles, which makes accurate sentiment identification challenging through conventional analytical tools. Consequently, there is a need for an efficient model capable of identifying sentiment directly from input text[9][10].

Accordingly, the primary problem addressed in this thesis is the difficulty of analyzing customer sentiment using traditional analytical approaches applied to varied textual data sources. The use of machine learning and deep learning methodologies can enhance sentiment classification and support organizational decision-making[11][12].

Thesis Objectives

1. Applying advanced deep neural network models to enhance customer feedback based on sentiment analysis
This objective aims to improve customer feedback analysis by applying advanced deep neural network models capable of capturing complex sentiment patterns.
2. Using benchmark dataset
A benchmark dataset is used to ensure objective evaluation and enable fair comparison with existing studies.
3. Evaluate model using different evaluation metrics
The models are evaluated using multiple metrics, including accuracy, precision, recall, and F1-score, to provide a comprehensive performance assessment.
4. Applying different preprocessing steps to improve text quality
Various text preprocessing techniques are applied to enhance data quality and improve sentiment classification performance.
5. Develop AI framework to enhance customer feedback analysis
An AI-based framework is developed to automate and support effective customer feedback analysis.
6. Conducted transformer model achieves the best performance compared with different machine learning models
Experimental results show that transformer-based models outperform traditional machine learning approaches in sentiment classification tasks.

Thesis Contribution:

This thesis provides a number of valuable contributions with regards to the existing literature on sentiment analysis as well as customer experience analytics. This thesis brings forth a comprehensive framework with regards to customer complaint sentiment analysis, combining sophisticated Natural Language Processing with machine learning, deep learning, as well as a transformer. This comprehensive framework encompasses a detailed preprocessing stage, which involves cleaning, normalization, as well as further processing through lemmatization, achieved with the use of WordNet.

This thesis further adds value through the comparative analysis of multiple models as different models such as Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, Support Vector Classifier, and BERT are applied to the sentiment classification task. The models will be tested on a benchmark dataset available on Kaggle, allowing for a thorough assessment of the models' robustness, as well as their adaptability capabilities. Additionally, this thesis incorporates comprehensive assessment metrics such as Accuracy, Precision, Recall, and F1 Score that would allow a better understanding of models' capabilities in recognizing dissatisfaction expressions through different paradigms of analysis, thus

providing valuable knowledge that may be practically applied through improved customer experience services.

Thesis Organization:

This thesis is organized into six chapters, structured according to the research flow and methodological framework.

Chapter 1: Introduction

This chapter presents the background and motivation of the thesis, followed by the problem definition, objectives, contributions, and the overall structure of the thesis.

Chapter 2: Overview

This chapter provides an introductory overview of the foundational concepts in sentiment analysis, Natural Language Processing (NLP), machine-learning techniques, deep-learning models, topic modeling, and transformer-based architectures. It establishes the theoretical background necessary for understanding the subsequent chapters.

Chapter 3: Related Work

This chapter presents an in-depth review of previous studies on sentiment analysis, covering classical ML models, DL models, BERT-based approaches, and complaint-specific sentiment research. It identifies research gaps that motivate the design of the proposed methodology.

Chapter 4: Proposed Approach

This chapter explains the complete methodological framework adopted in the thesis.

Chapter 5: Experimental Results and Discussion

This chapter presents the results obtained from evaluating all models on both datasets. It includes performance comparisons, confusion matrices, interpretation insights, and discussion of model strengths and limitations.

Chapter 6: Conclusion and Future Work

This chapter summarizes the key findings of the thesis and their implications for customer-experience analytics.

Chapter 2 Overview

Introduction

In this chapter, we present an overview of the main concepts related to this research, including the foundations of Artificial Intelligence, Machine Learning, Deep Learning, and Natural Language Processing, alongside their relevance to sentiment analysis in customer-complaint data.

2.1 Artificial Intelligence (AI)

Artificial Intelligence (AI) refers to computer systems that are capable of performing tasks that typically require human intelligence[13].

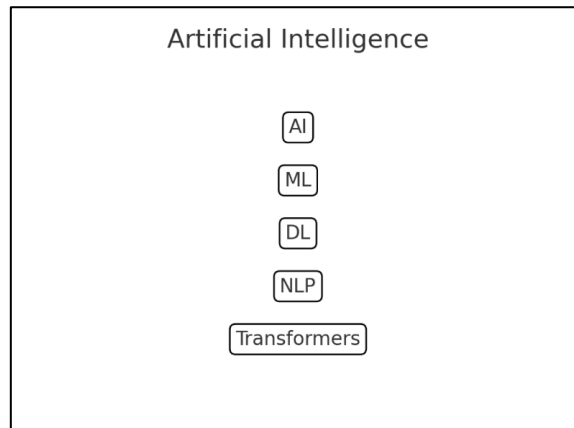


Fig. 2.1 – Artificial Intelligence (AI)

2.2 Natural Language Processing (NLP): Natural Language Processing (NLP) is a major subfield of AI concerned with enabling computers to understand, interpret, and generate human language[14].

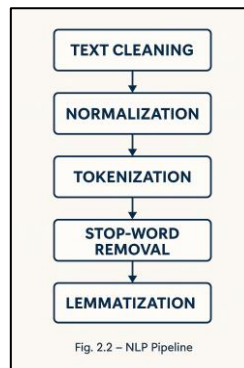


Fig. 2.2 – Natural Language Processing (NLP)

2.3 Machine Learning (ML): Machine Learning (ML), an application of AI, helps develop models that identify patterns in data using an algorithm that isn't rules-bound. ML works mostly on numeric values extracted from text, an example of this being TF-IDF[15].

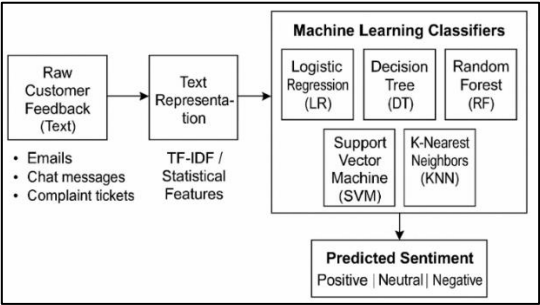


Fig. 2.3 – Machine Learning workflow

2.4 Deep Learning (DL): Deep Learning (DL) builds on the concept of ML and uses a type of neural network with multiple layers that enable the discovery of complex and abstract representations in the data[16].

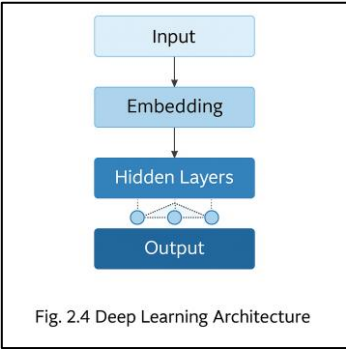


Fig. 2.4 Deep Learning Architecture

Fig. 2.4 – Deep Learning Architecture

2.5 Transformer Models (e.g., BERT): The transformer model, exemplified by BERT, represents a major advancement in the field of Natural Language Processing[17].

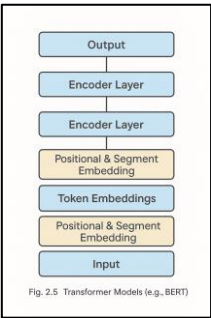


Fig. 2.5 Transformer Models (e.g., BERT)

Fig. 2.5 – Transformer (BERT) Architecture

2.6 Text Preprocessing Techniques[18]

- Cleaning
- Normalization
- Tokenization
- Stop Word Removal
- Lemmatization
- Label encoding

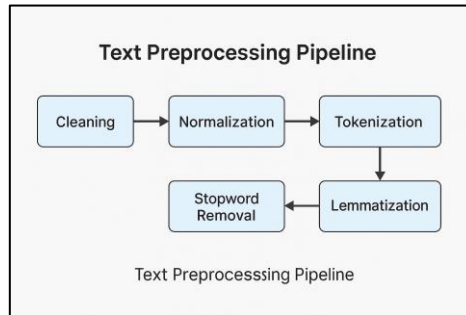


Fig. 2. 6 Text Preprocessing Techniques

2.7 Feature Engineering:

Feature engineering focuses on transforming raw text into informative signals that improve model performance. It includes generating n-grams, building vocabulary, and handling noisy or domain-specific linguistic patterns [19].

- N-grams
- Vocabulary Construction
- Handling Noisy Text

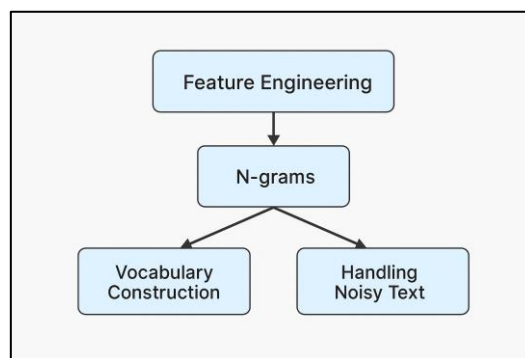


Fig. 2. 7 Feature Engineering

2.8 Feature Representation Techniques: Feature representation techniques transform unstructured text into numerical vectors that machine learning and deep learning models can understand. These representations capture different levels of linguistic information, ranging from simple word counts to advanced contextual meanings [20].

- Bag-of-Words (BoW)
- TF-IDF
- Word Embeddings
- Contextual Embeddings (BERT)

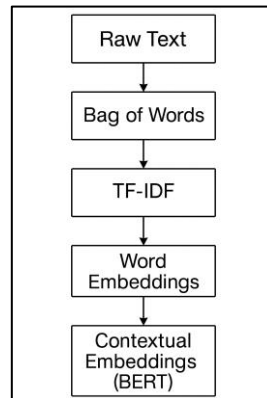


Fig. 2.8 Feature Representation Techniques

2.9 Classification Techniques: Classification is a type of supervised machine learning algorithm applied in the process of assigning labels like positive, negative, and neutral opinions in customer complaints [21].

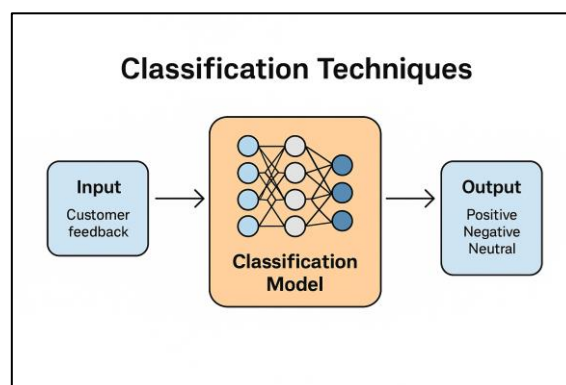


Fig. 2.9– Classification Techniques Workflow

2.9.1 Machine Learning Techniques

Decision Tree (DT): A Decision Tree is a rule-based classifier that splits text data using feature thresholds derived from numerical representations such as TF-IDF.

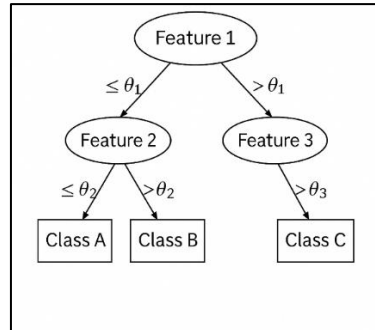


Fig. 2.9.1 - Decision Tree (DT)

Random Forest (RF):

Random Forest is an ensemble method that combines multiple Decision Trees trained on different subsets of data to reduce overfitting.

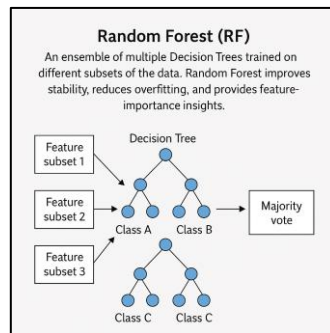


Fig. 2.9.1 - Random Forest (RF)

K-Nearest Neighbors (KNN):

The K-Nearest Neighbors algorithm, KNN, is an uncomplicated distance algorithm that classifies an input complaint's sentiment by analyzing the most frequent class among the K closest neighboring samples in the feature space.

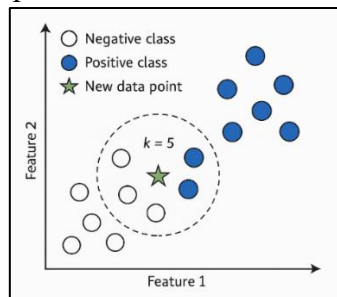


Fig. 2.9.1 - KNN

Support Vector Classifier (SVC):

The Support Vector Classifier (SVC) algorithm, a margin maximization algorithm, uses the concept of a hyperplane in a high-dimensional feature space. The SVC algorithm performs particularly well when the text data has TF-IDF feature mappings, because the algorithm works well with high-dimensional and sparse data.

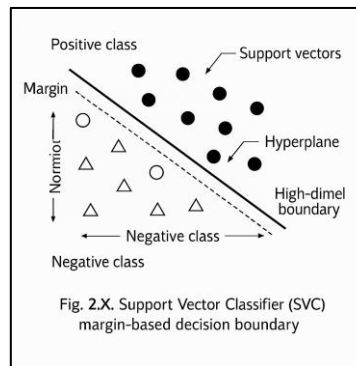


Fig. 2.9.1 – SVC

Naïve Bayes (NB):

A probabilistic machine learning algorithm based on Bayes' theorem that assumes independence between features. It is simple, fast, and commonly used for text classification tasks such as sentiment analysis and spam detection.

Extreme Gradient Boosting (XGBoost):

A powerful machine learning algorithm based on gradient boosting that is designed for high performance and scalability. It is widely used for classification and prediction tasks due to its ability to improve accuracy while reducing overfitting through regularization.

2.9.2 Deep Learning Techniques

Gated Recurrent Unit (GRU): Gated Recurrent Unit, commonly abbreviated as GRU, and Long Short-Term Memory, abbreviated as LSTM, are types of recurrent neural networks and play an important role in analyzing the sequential data present in the complaints registered by the customers.

Latent Class Topic Model (LCTM): Latent Class Topic Model, abbreviated as LCTM, has the characteristics of a probabilistic model. In this model, the topics in the complaints expressed by the customers are detected using the concept of grouping texts into latent topics.

Transformer-based Models (e.g., BERT): BERT, a transformer model, symbolizes the forefront of deep learning models in the field of NLP. Unlike other models like GRU and LSTMs, BERT

applies the concept of self-attention. BERT applies bidirectional training on the complaints lodged by the customers, resulting in better understanding and insight into the implicit emotions, services, and complaints.

DistilBERT:

A lightweight version of BERT that retains most of its language understanding capabilities while being smaller and faster. It is designed for efficient text classification tasks with reduced computational cost and minimal performance loss.

2.10 Feature Representation Techniques:

Feature representation is Representation in the feature space assumes great importance in the process of sentiment analysis, wherein unstructured text data gets converted into a numeric representation that can be understood and analyzed by machine and deep learning models. The process has largely involved TF-IDF, static word embeddings like Word2Vec and GloVe, and BERT, among others, and has also utilized the concept of topics, with models like LCTM [22].

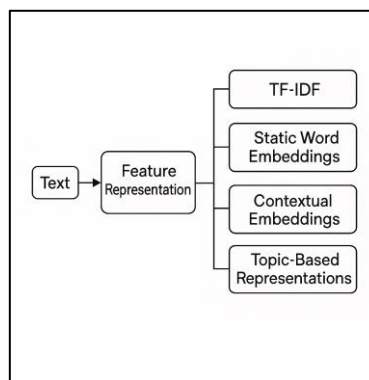


Fig. 2.10- Feature Representation

2.11 Evaluation Metrics: Metrics help in identifying the effectiveness of the model in classifying the unseen data. As the problem type here, in the case of the sentiment analysis problem, is classification, the performance would be measured in terms of the classification metrics.

- **Accuracy**
- **Precision**
- **Recall**
- **F1-Score**
- **Confusion matrix**

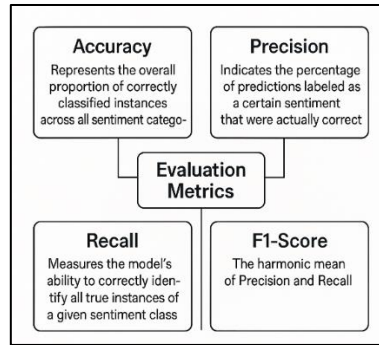


Fig. 2.11 - Evaluation Metrics

2.12 Sentiment Analysis: Concepts and Categories: Sentiment analysis is a core Natural Language Processing task focused on identifying the emotional orientation expressed within text. In customer-complaint environments, sentiment analysis classifies feedback into categories such as positive, negative, and neutral, enabling organizations to understand customer emotions, detect dissatisfaction triggers, and extract actionable insights from large volumes of unstructured text. This process supports data-driven decision-making and improves the overall quality of customer-experience management [23].

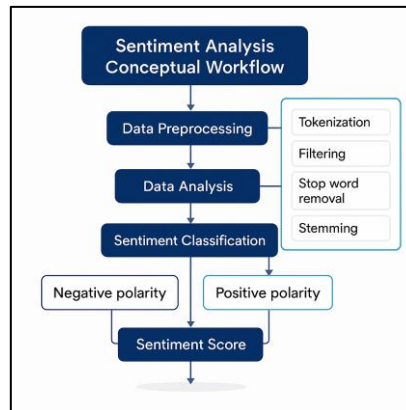


Fig. 2.11 – Sentiment Analysis Conceptual Workflow

Chapter 3 Related Work

The need for sentiment analysis in customer analysis has been driven by the popularity of user-generated content on platforms like Amazon and Twitter, as well as on airline review and e-commerce sites. An increase in the number of customer reviews online has also led to the application of machine learning, deep learning, and transformers in studies of customer analytics. This section sheds light on key studies and the gaps in customer feedback analysis.

The study[24] analyzed consumer sentiment toward a telecommunication operator's responses using Twitter data. The dataset used comprised consumer feedback in the form of tweets about telecommunication services. Preprocessing involved text normalization to remove social media noisy text, abbreviations, and language used in the text. However, the proposed system used Deep Convolutional Neural Networks (DCNN) to extract features, along with an SVM classifier. The study's evaluation yielded 63% accuracy, 63% precision, and 50% recall, indicating that the system lacked the ability to retrieve accurate instances of the sentiment. Actually, the authors noted that the size of the posts, as well as their variability in language, pose challenges to the recall levels, specifically for the minority classes, in this case, the sentiments.

In the study[25], the task was sentiment classification on highly noisy consumer feedback data collected by Global Support Services and Knowledge Base surveys. There were over 40,000 instances in the dataset, each linked with satisfaction levels ranging from “not satisfied” to “very satisfied.” In the preprocessing phase, lemmatization, stop-word removal, and feature reduction were performed. The linear Support Vector Machine approach using surface-level n-grams combined with linguistic features was used. In this experiment, it was shown that performance was moderate, with an F1-score of 74.62% at extreme levels of satisfaction, but it sharply decreased to 58.14% for highly related classes of consumer sentiment. These experiments show the challenges encountered in obtaining high recall on noisy consumer feedback on sentiment boundaries.

The study[26] the work was a part of the IJCNLP-2017 Shared Task on Customer Feedback Analysis, and the task was to classify the multilingual customer feedback from Microsoft Office users. The experiment involved four languages—English, French, Spanish, and Japanese—and predefined categories included feedback such as complaint, request, and others. The machine learning techniques used were Support Vector Machine (SVM), the Naive Bayes Classifier, and the bidirectional LSTM model with word embeddings. According to the results, the exact match accuracy on English feedback was around 65%, while the least was 56% for Japanese feedback due to the difficulty in identifying sentiment-related feedback.

The study[27] proposes and examines a dataset for customer feedback analysis in the context of the IJCNLP-2017 Shared Task, which contained feedback in four languages for six semantic categories. The dataset tested systems for their exact match accuracy, micro-averaged precision, recall, and F1-measure; although some systems delivered reasonable levels of accuracy, it was found that there was considerable difficulty for models in accessing all possible tags, especially for multi-label tasks, and because macro-averaged recalls were significantly impacted due to extreme class imbalance, which was challenging for models to identify complaint-related feedback text correctly. This clearly indicates difficulties associated with existing techniques in practical applications for analyzing customer feedback.

The study[28] focused on analyzing customer reviews on the web using contemporary natural language processing techniques, including BERT models and multilingual zero-shot classifiers. The dataset used for this experiment was taken from customer reviews on restaurant sites for customers who speak Turkish. The preprocessing steps included normalization suited to the structure of an agglutinative language. Despite using sophisticated models, this study found variability in recall across sentiment categories, especially for the negative class.

In the study[29], the authors proposed the use of a Convolutional Neural Network (CNN) and a Gated Recurrent Unit (GRU)- based automated system for customer feedback analysis. The experiment used multilingual customer feedback datasets contributed to the IJCNLP shared task on customer feedback analysis, including datasets for English, French, Spanish, and Japanese. The authors of the research work proposed using deep learning models to reduce the manual feature engineering process. The results of the experiments showed that the models produced a moderate level of accuracy of 71.75% for the French dataset, while the accuracy of the remaining datasets was lower; however, the authors highlighted that the customer feedback analysis for multilingual datasets is a complex task, and hence the recall level of the models was restricted.

Chapter 4 Proposed Approach

4.1 Introduction:

This chapter describes the proposed conceptual framework that was designed for the automatic sentiment classification of customer complaints recorded in the public dataset.

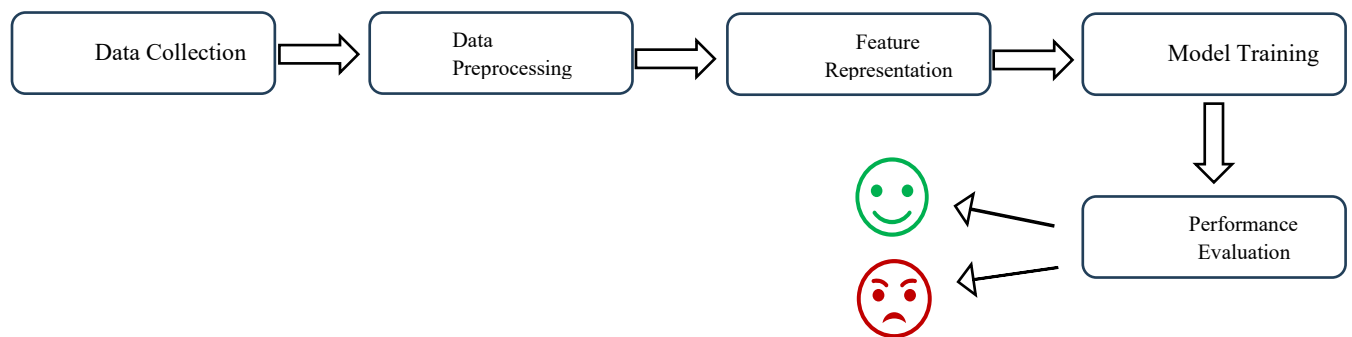
This chapter helps in comprehending the system architecture as a whole, right from gathering data, preprocessing, feature development, modeling, as well as the evaluation process. The proposed system incorporates TF-IDF Extracted Features, together with word count Extracted Features from BERT, for better identification of negative as well as neutral complain statements in real-time[23].

4.2 Problem Context and Rationale: Customer dissatisfaction often triggers escalation much before any formal complaint is filed. Customer complaints involve extensive unstructured text, thereby making text analysis cumbersome. This gives rise to problems such as:

- Late identification of high-risk negative complaints
- Downloading overloaded ticketing
- Hidden negative sentiment inside lengthy text descriptions
- Lack of visibility of dissatisfaction trends

An automated framework that performs sentiment analysis assists in the early identification of important issues, better routing, and improved customer sentiment analysis.

Fig. 4 – Methodology



A. Data Collection

- Data was collected from **Kaggle dataset**: Customer Complaints Sentiment and Priority Dataset
Example entries include:
- The dataset represents real-world customer feedback characterized by diverse linguistic structures, varying complaint lengths, and implicit expressions of dissatisfaction.
- The input feature is the Consumer_complaint text field, and the target variable Sentiment is encoded as a binary label: 0 for non-negative sentiment and 1 for negative sentiment.

Dataset Structure:**Complaint Text:**

Free-text customer complaint descriptions expressing dissatisfaction and service-related issues.

Sentiment Label

0 – Negative

1 – Positive

Dataset Characteristics:

- **Data type:** Structured textual data
- **Language:** English
- **Task:** Binary sentiment classification
- **Domain:** Customer complaints
- **Class distribution:** Imbalanced

Feature Representation Techniques:**Class Distribution:**

- **Total samples:** 1,750
- **Positive complaints:** 1,295 (74%)
- **Negative complaints:** 455 (26%)

Example Complaint (Customer complaint)	Sentiment	Priority
<i>"I had overdraft protection... however instead..."</i>	0	1
<i>"US Bank rearranged transactions causing multiple overdrafts..."</i>	1	1
<i>"Hello, this email is in reference to the fraud department..."</i>	0	0

Table 4.A – Data Example

B. Data Preprocessing

The data preprocessing stage focused on improving text quality and ensuring consistency before model training.

- **Text**

Cleaning:

Raw complaint texts were cleaned by removing noise such as special characters, punctuation, and unnecessary symbols to improve readability and model performance.

- **Text**

Normalization:

Stop-word removal and stemming were applied only for traditional feature extraction methods such as Word Count and TF-IDF, in order to reduce dimensionality and standardize word forms.

- **Label** **Formatting:**
Sentiment labels were converted into a numerical format to ensure compatibility with machine learning and deep learning models.

Data Preparation

This stage ensured that the data was ready for experimentation and fair evaluation.

- **Experimental** **Environment:**
All experiments were implemented using Jupyter Notebook on a Windows 10 machine with an Intel Core i7 CPU and 16 GB RAM, utilizing Python and standard libraries such as pandas and scikit-learn.
- **Train/Test** **Split:**
The dataset was divided into training and testing sets to evaluate model performance on unseen data.
- **Imbalanced** **Class** **Distribution:**
The dataset showed class imbalance, with 1,295 negative samples (74%) and 455 positive samples (26%), which was considered during model evaluation and performance analysis.

Dataset	Sets	Positive (0)	Negative (1)	Total
Customer Complaints	Training set	1,036	364	1,400
	Testing set	259	91	350
Total		1,295	455	1,750

Table 4.B – Data Preparation

C. Feature Representation Techniques

Word Count Representation

The word count technique was used to transform complaint texts into numerical vectors based on raw term frequency. This representation serves as a baseline approach and captures word occurrence patterns without considering contextual dependencies.

TF-IDF Representation

Term Frequency–Inverse Document Frequency (TF-IDF) was employed to weight words according to their importance across the corpus. This technique reduces the influence of frequent but non-informative words while highlighting discriminative terms relevant to sentiment classification.

D. Model Training:

Machine Learning Models

The following machine learning classifiers were implemented and evaluated using both word count and TF-IDF representations:

- Logistic Regression (LR)
- Random Forest (RF)
- Decision Tree (DT)
- Naïve Bayes (NB)
- K-Nearest Neighbors (KNN)
- Extreme Gradient Boosting (XGBoost)

These models were selected due to their robustness, interpretability, and widespread use in text classification research.

Transformer-Based Models

To capture deeper semantic and contextual information, two transformer-based models were fine-tuned:

- **BERT (Bidirectional Encoder Representations from Transformers)**
- **DistilBERT**, a compressed version of BERT designed to reduce computational complexity while maintaining performance

Both models were initialized using pretrained weights and fine-tuned on the complaint dataset using supervised learning.

E. Evaluation Metrics

Metric	Description	Evaluation Formula
Accuracy	Measures overall correctness of predictions	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$
Precision	Measures correctness of predicted positive class	$Precision = \frac{TP}{TP + FP}$
Recall	Measures ability to detect all actual positive cases	$Recall = \frac{TP}{TP + FN}$
F1-score	Harmonic mean of precision and recall	$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$

Table 4.E – Evaluation Metrics

Confusion Matrix Structure:

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Table 4.E – Confusion Metrics

Chapter 5 Experimental Results and Discussion

5.1 Introduction

This chapter will introduce and interpret the results of applying a variety of sentiment classification models to the customer complaints dataset. The purpose of this chapter is to compare traditional machine learning models with deep learning models based on different feature representation methods.

In this thesis, the use of standard performance measures such as Accuracy, Precision, Recall, F1 score, and confusion matrix enables a comprehensive comparison. Special focus will be placed on the influence of feature representation, imbalanced classes, and the capacity of models to identify effective, tacitly patterned consumer dissatisfaction that usually permeates consumer complaints.

5.2 Results of Machine Learning Models Using Word Count

Based on the Word Count figure, the results from machine learning are moderate. The highest accuracy was achieved by the Random Forest (0.7543) and XGBoost (0.7371) models, both of which are ensemble learning methods that perform well on simpler features such as word count.

Logistic Regression and Naïve Bayes performed tolerably but were inferior, whereas Decision Trees suffered from overfitting. The worst performance was observed with K-Nearest Neighbors, with an accuracy of 0.5057, indicating sensitivity to sparse, high-dimensional features in text data.

Essentially, the Word Count functions' ability to identify meaning, particularly the complex, implied meaning of customer complaints, was limited.

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression (LR)	0.6914	0.6545	0.6914	0.6675
Random Forest (RF)	0.7543	0.7222	0.7543	0.7012
Decision Tree (DT)	0.6200	0.6240	0.6200	0.6220
Naïve Bayes (NB)	0.6629	0.6320	0.6629	0.6447
K-Nearest Neighbors (KNN)	0.5057	0.7188	0.5057	0.5221
XGBoost (XGB)	0.7371	0.6968	0.7371	0.6989

Table 5.2 – ML Using Word Count Results

5.2.1 Confusion Matrix Analysis – Word Count Models

From the confusion matrices, most algorithms in the Word Count category tended to classify the majority class. This resulted in a higher number of false negatives, especially for the minority class expressing negative sentiment. K-NN was the poorest performer in terms of misclassification error. Notably, the remaining algorithms performed well but had a large margin of error. This upholds the ineffectiveness of the frequency-based attribute.

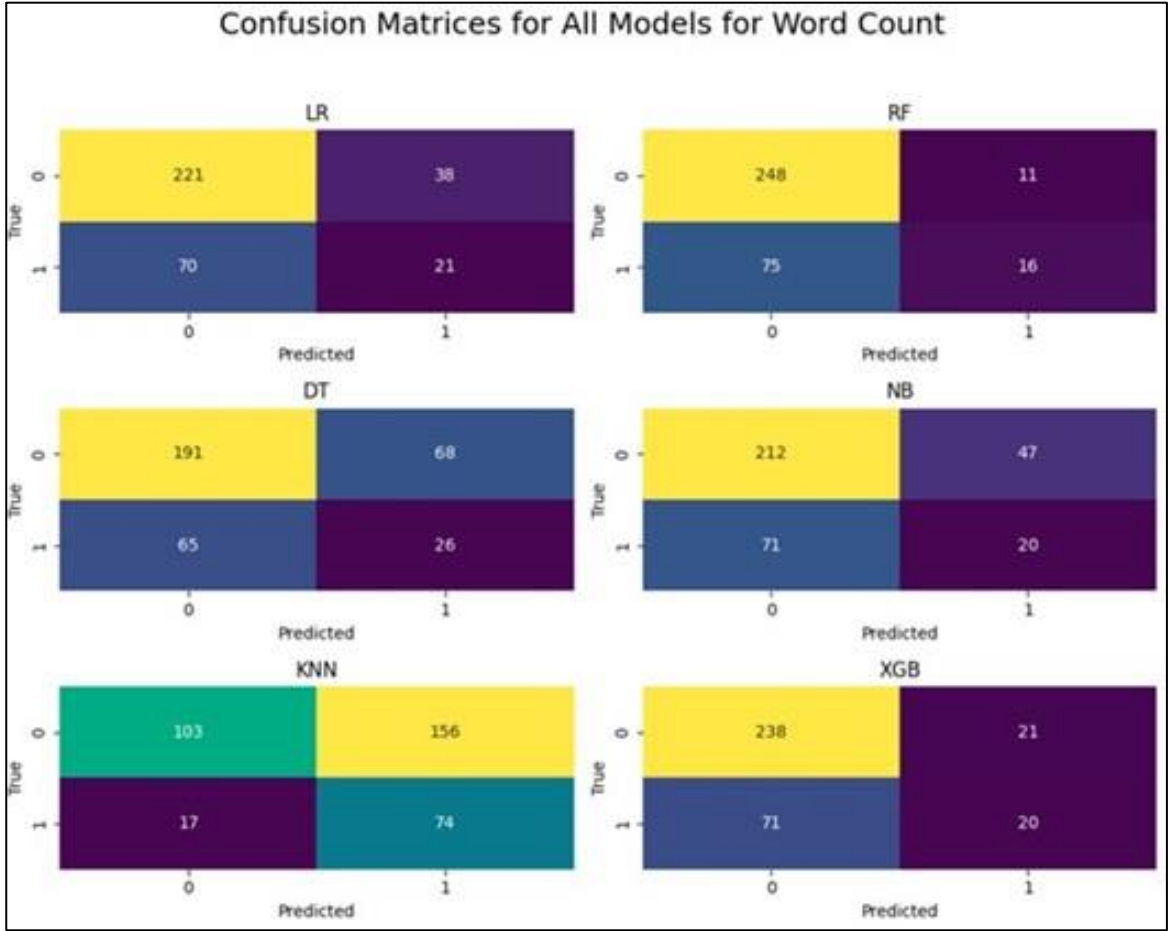


Fig.5.2.1 Confusion Matrix- Word Count Models

5.3 Results of Machine Learning Models Using TF-IDF

The TF-IDF representation showed improved performance over Word Count across most machine learning models. The best-performing classical model was XGBoost with an accuracy of 0.7685 and an F1 score of 0.7379.

Random Forest and Naïve Bayes were also aided by TF-IDF, as it gave higher weight to discriminative features while diminishing the impact of general terms. Yet, despite this, some models showed inequity between precision and recall, indicating the difficulties in correctly identifying minority sentiment categories.

TF-IDF appeared to be more effective than Word Counting, but its vector form has the disadvantage that there are very few non-zero entries.

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression (LR)	0.7057	0.6612	0.7057	0.6737
Random Forest (RF)	0.7371	0.6615	0.7371	0.6428
Decision Tree (DT)	0.6457	0.6507	0.6457	0.6481
Naïve Bayes (NB)	0.7400	0.5476	0.7400	0.6294
K-Nearest Neighbors (KNN)	0.6942	0.6428	0.6942	0.6587
XGBoost (XGB)	0.7685	0.7447	0.7685	0.7379

Table 5.3 – ML Using TF-IDF Results

5.3.1 Confusion Matrix Analysis – TF-IDF Models

The confusion matrices of models developed using the TF-IDF technique showed better class separation than those of the Word Count method. XGBoost and Random Forest showed lower false-positive rates and better negative-sample detection. Yet Naïve Bayes and Decision Tree still struggled to correctly classify a large number of data points in the minority class, suggesting that traditional models are not good at expressing sentiment.

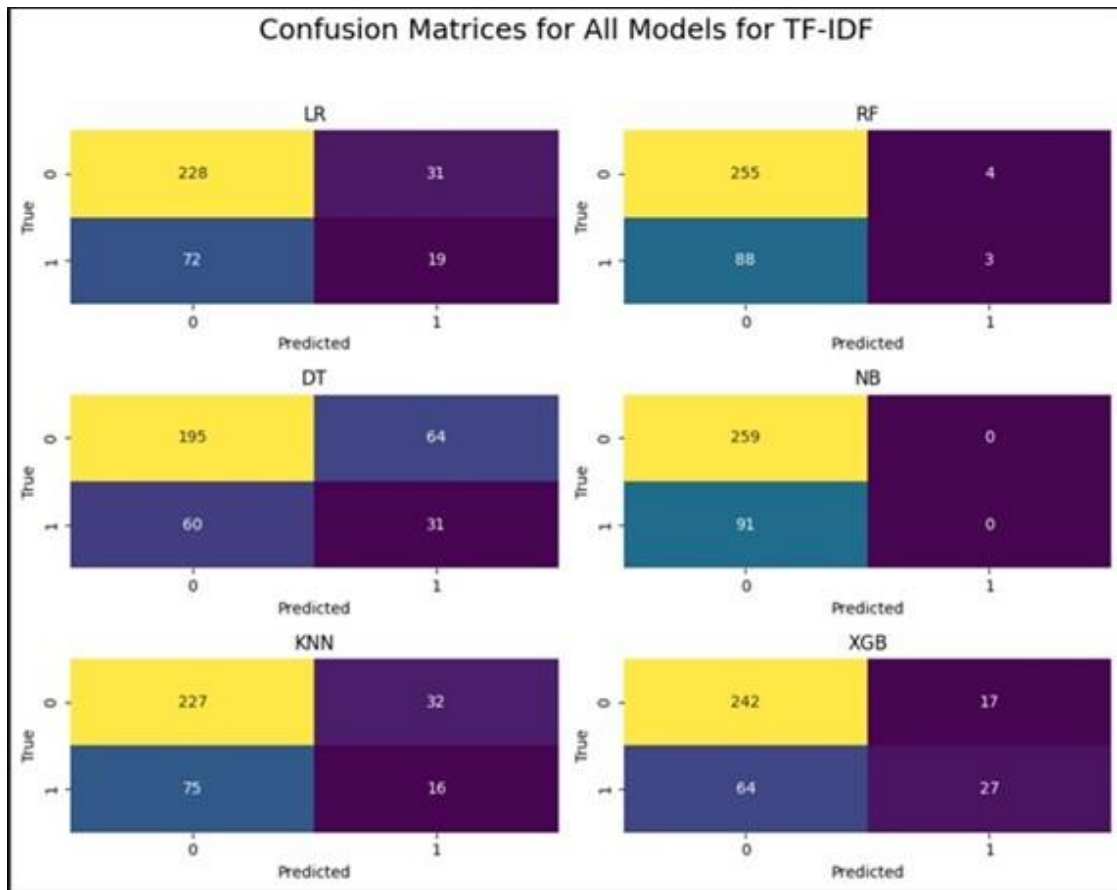


Fig.5.3.1 Confusion Matrix- TF-IDF Models

5.4 Results of Transformer-Based Models

Transformer models showed excellent and consistent performance compared to traditional machine learning models. BERT achieved the best overall accuracy (0.7486) and F1-score (0.7322) among the models, and DistilBERT showed similar results with slightly lower values. Still, the fact that DistilBERT is comparatively compact allowed it to retain the core ability to provide the same level of understanding in the context as BERT, making it a highly feasible choice in terms of performance and processing cost.

Model	Accuracy	Precision	Recall	F1-score
BERT	0.7486	0.7266	0.7486	0.7322
DistilBERT	0.7457	0.7199	0.7457	0.7252

Table 5.4 – Transformer Based Models Results

5.4.1 Confusion Matrix Analysis – Transformer Models

The confusion matrices for BERT and DistilBERT are found to have a well-proportioned distribution of true and false instances, compared with the confusion matrices for other models. The results reveal that for BERT and DistilBERT, the misclassification rate for the minority sentiment class has decreased dramatically.

BERT showed marginally better performance on measures of class discrimination, although DistilBERT was once again a close rival in terms of computational cost. The rest were mostly related to pending or borderline cases.

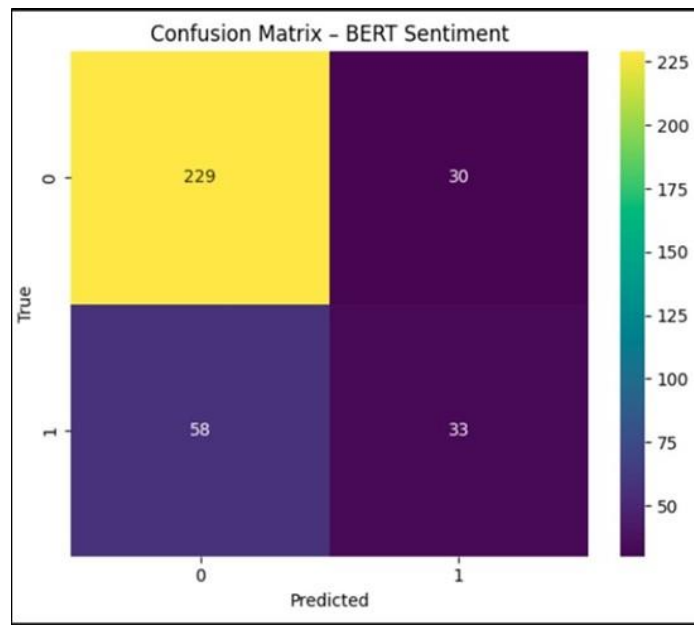


Fig.5.4.1 Confusion Matrix- Transformer Models BERT

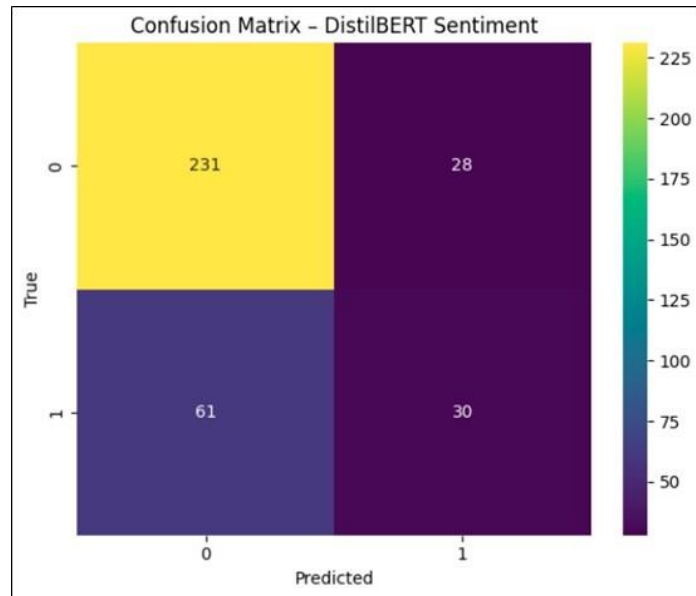


Fig.5.4.1 Confusion Matrix- Transformer Models DistilBERT

5.5 Overall Comparison and Discussion

The experimental results lead to several key findings:

- TF-IDF outperforms Word Count for traditional machine learning models.
- XGBoost is the strongest classical machine learning model across both feature representations.
- Transformer-based models outperform most traditional approaches in terms of balanced performance.
- DistilBERT provides an effective balance between accuracy and efficiency.
- Context-aware embeddings are critical for detecting implicit dissatisfaction in customer complaints.

These findings confirm that transformer-based models are better suited to real-world customer experience applications, where textual data is complex, imbalanced, and context-dependent.

Chapter 6 Conclusion and Future Work

6.1 Introduction

This final chapter summarizes the findings from the different sentiment analysis tasks conducted and the insights that can be extracted to enhance the customer experience, in addition to outlining possible areas for improvement in advancing the proposed methods or techniques, which justify why this study is worth pursuing.

The focus of this chapter is to highlight the need for advanced applications of machine learning and transformer models to enable predictive decisions and proactive identification of dissatisfied customers.

6.2 Conclusion

This thesis, the performance of traditional machine learning algorithms and deep learning techniques based on the transformer architecture was evaluated for sentiment analysis of customer complaints. It was found that feature presentation remains crucial, and TF-IDF proved more effective than Word Count across most machine learning models.

In classical approaches, the best-performing model was XGBoost, demonstrating the applicability of ensemble learning to structured text features. However, transformer models like BERT and DistilBERT performed comparatively better by understanding the context in written complaint forms.

The outcome highlights the limitations of traditional machine learning algorithms in handling implicit sentiment expression, thereby underscoring the usefulness of transformer models for realistic customer feedback analysis. Nevertheless, the proposed method provides a strong foundation for data-driven customer-experience strategy formulation based on sentiment analysis.

6.3 Future Work

Several directions can be explored to extend and enhance the proposed sentiment analysis framework:

Some ways in which the proposed framework for sentiment analysis can be developed and advanced are:

- **Early Sentiment Analysis:**
Future work would then include optimizing the transformer model to detect negative customer sentiment earlier, enabling proactive measures to be taken before escalation.
- **Real-Time CX Monitoring:**
Moreover, sentiment analysis can be implemented in real-time dashboards for quick decision-making in businesses.
- **Expanded Data Coverage:**
Incorporating more types of data, such as social networking site touchpoints, live conversations, or call center transcripts, into the existing data in a review might help better understand consumer sentiment.
- **Continuous CX improvement:** Modeling techniques used in real applications can help improve and monitor performance, as well as estimate business outcomes.

References

- [1] B. Liu, "Sentiment Analysis and Opinion Mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, May 2012, doi: 10.2200/S00416ED1V01Y201204HLT016.
- [2] S. Hossain and S. Tarake, "Comparative Analysis of Sentiment Analysis Models for Consumer Feedback: Evaluating the Impact of Machine Learning and Deep Learning Approaches on Business Strategies," pp. 18–29, 2025.
- [3] M. Rafiq, M. Yousaf, Y. Aziz, M. Fuzail, A. Naeem, and N. Aslam, "SENTIMENT ANALYSIS OF CUSTOMER FEEDBACK ON ONLINE PLATFORM USING DEEP LEARNING," *Kashf Journal of Multidisciplinary Research*, pp. 2–06, 2025, Accessed: Dec. 02, 2025. [Online]. Available: <https://kjmr.com.pk>
- [4] A. Mohammed and R. Kora, "Deep Learning Approaches for Arabic Sentiment Analysis," *Soc Netw Anal Min*, vol. 9, no. 1, 2019.
- [5] A. Patel, P. Oza, and S. Agrawal, "Sentiment Analysis of Customer Feedback and Reviews for Airline Services using Language Representation Model," *Procedia Comput Sci*, pp. 2459–2467, 2022.
- [6] xJourney, "Customer Complaints Sentiment and Priority Dataset," 2023. Accessed: Dec. 02, 2025. [Online]. Available: : <https://www.kaggle.com/datasets/xjoury/customer-complaints-sentiment-and-priority-dataset>
- [7] C. Fan, M. Jiang, and Y. Wu, "Early Detection of Customer Dissatisfaction Using Text Analytics," *Inf Process Manag*, vol. 58, no. 3, 2021, Accessed: Dec. 02, 2025. [Online]. Available: 10.1016/j.ipm.2021.102540
- [8] L. Kulkarni and S. Roy, "Real-Time Sentiment Monitoring for Customer Feedback Streams," *Expert Syst Appl*, vol. 189, 2022.
- [9] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning-Based Text Classification: A Comprehensive Review," *ACM Comput Surv*, vol. 54, no. 3, 2021.
- [10] S. Zhang, X. Wang, and H. Chen, "Handling Linguistic Variability in Customer Feedback Classification," *Knowl Based Syst*, vol. 234, 2021.
- [11] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.
- [12] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in *EMNLP*, 2014.
- [13] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 4th Edition. 2021. Accessed: Dec. 02, 2025. [Online]. Available: <https://aima.cs.berkeley.edu/>
- [14] Daniel Jurafsky and James H. Martin, *Speech and Language Processing*, 3rd Edition (Draft). Stanford University, 2023.
- [15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Edition. Springer, 2009. doi: <https://doi.org/10.1007/978-0-387-84858-7>.
- [16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*. MIT Press, 2016. Accessed: Dec. 02, 2025. [Online]. Available: <https://www.deeplearningbook.org/>
- [17] Steven Bird, Ewan Klein, and Edward Loper, *Natural Language Processing with Python*. O'Reilly Media, 2009. Accessed: Dec. 02, 2025. [Online]. Available: <https://www.nltk.org/book/>
- [18] Alice Zheng and Amanda Casari, *Feature Engineering for Machine Learning: Principles and Techniques*. O'Reilly Media, 2018. Accessed: Dec. 02, 2025. [Online]. Available: <https://www.oreilly.com/library/view/feature-engineering-for/9781491953235/>
- [19] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd Edition (Draft). Pearson / Stanford University, 2025. Accessed: Dec. 02, 2025. [Online]. Available: <https://web.stanford.edu/~jurafsky/slp3/>
- [20] Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd Edition. 2019. Accessed: Dec. 02, 2025. [Online]. Available: <https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/>
- [21] Daniel Jurafsky and James H. Martin, *Speech and Language Processing*, 3rd Edition (Draft). Pearson / Stanford University, 2025. Accessed: Dec. 02, 2025. [Online]. Available: <https://web.stanford.edu/~jurafsky/slp3/>
- [22] Bing Liu, *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012. Accessed: Dec. 02, 2025. [Online]. Available: <https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf>
- [23] Wafa Alosaimi, Sara Alotaibi, Abeer Alsubaihini, and Rawan Alshammari, "ArabBERT-LSTM Improving Arabic Sentiment Analysis Based on Transformer Model and Long Short-Term Memory," *Front Artif Intell*, 2024, Accessed: Dec. 02, 2025. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/frai.2024.1408845>
- [24] S. Hossain *et al.*, "Comparative Analysis of Sentiment Analysis Models for Consumer Feedback: Evaluating the Impact of Machine Learning and Deep Learning Approaches on Business Strategies," vol. 05, pp. 18–29, 2025, doi: 10.37547/marketing.
- [25] M. Rafiq, M2. Yousaf, Y. Aziz, M. Fuzail, A. Naeem, and N. Aslam, "SENTIMENT ANALYSIS OF CUSTOMER FEEDBACK ON ONLINE PLATFORM USING DEEP LEARNING," *Kashf Journal of Multidisciplinary Research*, pp. 2–06, 2025, [Online]. Available: <https://kjmr.com.pk>
- [26] A. Mohammed and R. Kora, "Deep learning approaches for Arabic sentiment analysis," *Soc Netw Anal Min*, vol. 9, no. 1, Dec. 2019, doi: 10.1007/s13278-019-0596-4.

- [27] A. Patel, P. Oza, and S. Agrawal, "Sentiment Analysis of Customer Feedback and Reviews for Airline Services using Language Representation Model," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 2459–2467. doi: 10.1016/j.procs.2023.01.221.
- [28] H. Almaqtari, F. Zeng, and A. Mohammed, "Enhancing Arabic Sentiment Analysis of Consumer Reviews: Machine Learning and Deep Learning Methods Based on NLP," *Algorithms*, vol. 17, no. 11, Nov. 2024, doi: 10.3390/a17110495.
- [29] A. Mohammad Salloum and M. MAZEN Almustafa, "ANALYSIS AND CLASSIFICATION OF CUSTOMER REVIEWS IN ARABIC USING MACHINE LEARNING AND DEEP LEARNING," 2023, doi: 10.5281/zenodo.777803.