



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 increasing adoption of customer experience analytics practices includes the aspect of sentiment analysis involved in the decoding process. Rapid growth in terms of digital forms of communication contact, such as the use of chat services, emails, as well as service websites, has contributed to the increasing demand for a more efficient model in the classification of service delivery sentiment.

In this thesis, a comprehensive method for sentiment classification of customer complaints has been designed through the integration of traditional machine models with the use of transformers based on the idea of deep learning. This method will be tested on the public dataset 'Customer_Complaints_Sentiment' present on Kaggle, Customer complaints can be found in text form with labels of positive, negative, or neutral.

The whole preprocessing module within the conceptual viewpoint of the Natural Language Processing (NLP) task has been designed, which includes every activity that is being presented in this thesis. Once the preprocessing task is carried out, the models, such as Logistic Regression (LR), Decision Trees (DT), Linear Support Vector Classifier (Linear SVC), and others, such as K-Nearest Neighbors (KNN), BERT, were designed, along with the performance measures, such as Accuracy, Precision, Recall, F1 Score, and ROC-AUC.

Experimental results show that there is a difference in the level of performance of the models. Conventional machine learning models such as Decision Tree, KNN, and Linear SVC had a low level of efficiency, especially because of their low recall value and inability to identify negative sentiment expressions. Logistic Regression gave accuracy of 0.7448, but its recall value was low (0.0221). Consequently, its F1-score was 0.0429. Decision Tree gave accuracy of 0.6324 with F1-score of 0.3545. KNN gave accuracy of 0.7390 with F1-score of 0.1697. The accuracy of the well-tuned Linear SVC was 0.7467, but its recall value was low (0.0956) with F1-score of 0.1635.

The best-performing model was that of BERT with the decision threshold value of 0.45, which gave the perfect trade-off between the performance (Accuracy = 76%, F1 Score = 0.44) measures with significantly improved negative complaint detection. These results show that the transformer models clearly surpass the traditional models in addressing domain-specific patient complaints expressed through narratives. What this work adds in terms of value is that, unlike other studies, this work gives a comprehensive analysis involving multiple models concerning customer complaints within an organization, which still remains relatively unaddressed in existing literature compared to studies related to generic review datasets. This methodology will be helpful with its insights for any organization planning to adapt the use of sentiment analysis.

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	22
A. Data Collection.....	23
B. Data Preprocessing	24
C. Feature Engineering	24
D. Feature Representation Techniques	25
E. Classification Techniques	25
F. Feature Representation Techniques	26
Chapter 6 Conclusion and Future Work.....	34
Conclusion:	34
6.1 Important Features	34
6.3 Future Work	34
References	35

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 – The main steps of the Sentiment Analysis Framework
Figure 5.2.1: Comparative Performance Discussion
Figure 5.3 - Performance Comparison of the Best Models on the Customer_Complaints Dataset
Figure 5.4.1: Confusion Matrix for BERT
Figure 5.4.2 — Precision–Recall Curves for All Models

List of Tables

Table 4.3.A – Example of customer complaint.....	23
Table 4.3.E – Evaluation Metrics.....	24
Table 4.3.E.2 – Confusion Matrix Structure.....	24
Table 5.2 Overall Model Performance.....	25

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 **Develop an integrated sentiment-analysis framework for customer complaints**
To establish a unified and automated approach for analyzing customer complaints, enabling organizations to systematically understand sentiment patterns and improve customer-experience decision-making.
- 2 **Apply advanced preprocessing and linguistic-enhancement techniques to improve data quality and understand customer behavior**
To implement multiple text-preprocessing strategies that reduce noise, improve linguistic clarity, and support deeper insights into customer behavior and service-quality issues.
- 3 **Evaluate the impact of different deep-learning strategies on sentiment-classification performance**
To investigate how modern deep-learning approaches enhance the accuracy, robustness, and generalization capability of sentiment-analysis systems.
- 4 **Employ transformer-based architectures and comprehensive evaluation metrics for performance benchmarking**
To examine the contribution of state-of-the-art transformer models within the sentiment-analysis workflow and evaluate their performance using rigorous classification metrics.
- 5 **Enhance service quality through insights extracted from customer complaints**
To identify recurring patterns and issues within customer complaints that help organizations understand the sources of dissatisfaction and improve service quality.

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].

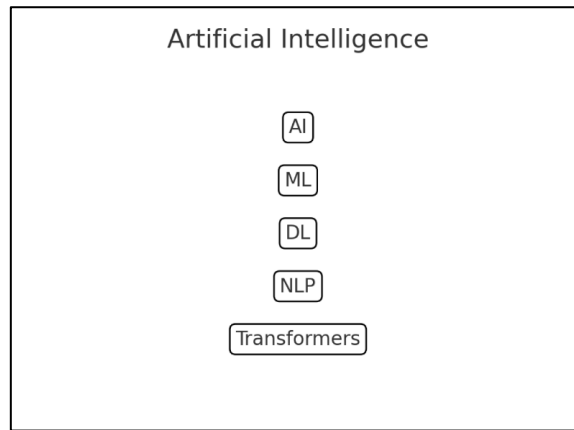


Fig. 2.1 – Artificial Intelligence (AI)

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].

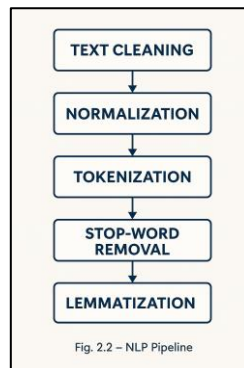


Fig. 2.2 – Natural Language Processing (NLP)

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].

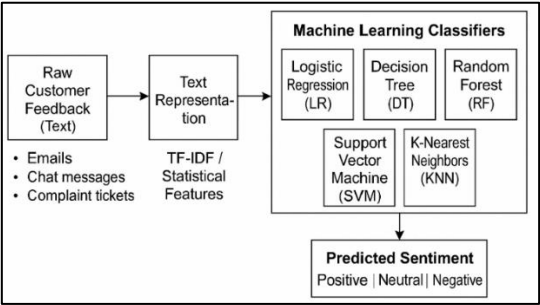


Fig. 2.3 – Machine Learning workflow

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].

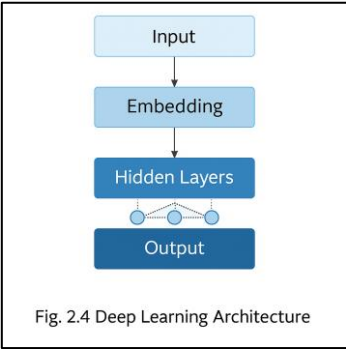


Fig. 2.4 Deep Learning Architecture

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].

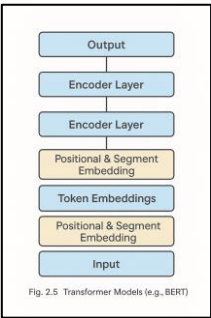


Fig. 2.5 Transformer Models (e.g., BERT)

Fig. 2.5 – Transformer (BERT) Architecture

2.6 Text Preprocessing Techniques[18]

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

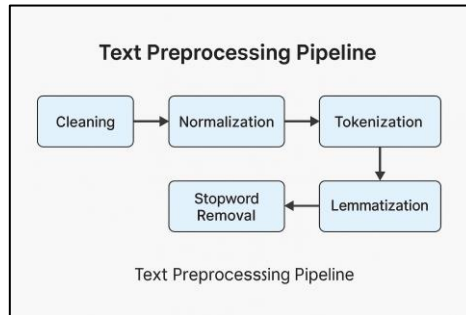


Fig. 2. 6 Text Preprocessing Techniques

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

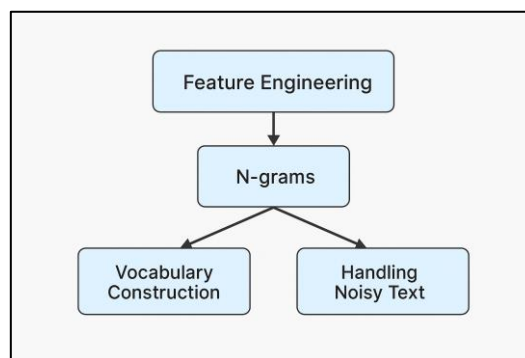


Fig. 2. 7 Feature Engineering

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)

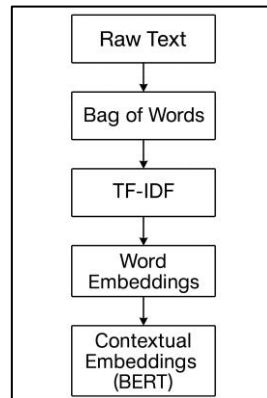


Fig. 2.8 Feature Representation Techniques

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].

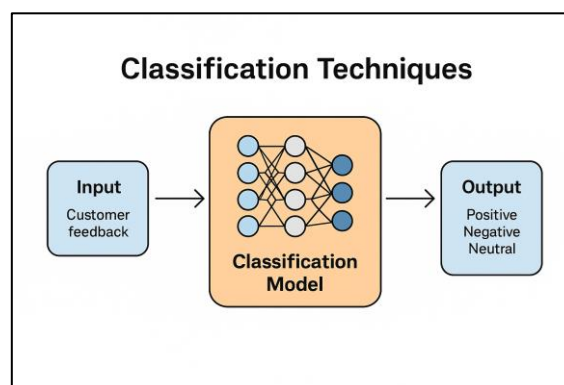


Fig. 2.9– Classification Techniques Workflow

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.

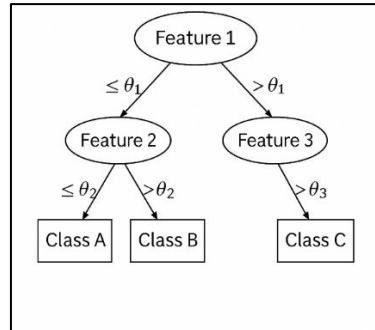


Fig. 2.9.1 - Decision Tree (DT)

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

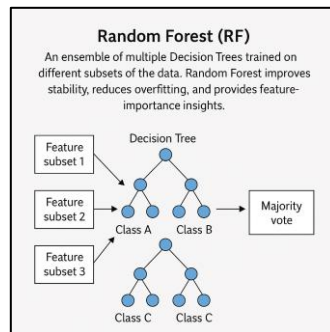


Fig. 2.9.1 - Random Forest (RF)

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.

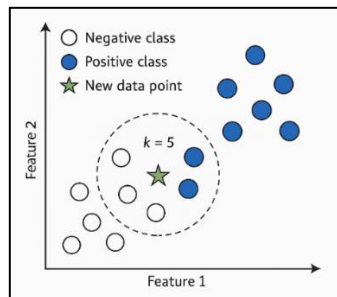


Fig. 2.9.1 - KNN

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.

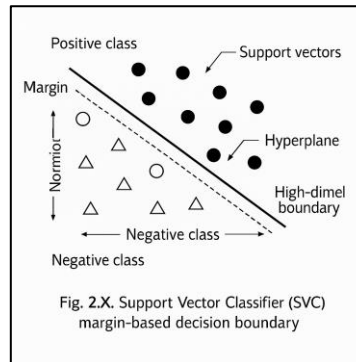


Fig. 2.9.1 – SVC

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.

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].

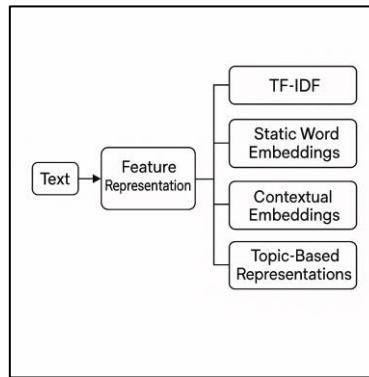


Fig. 2.10- Feature Representation

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**

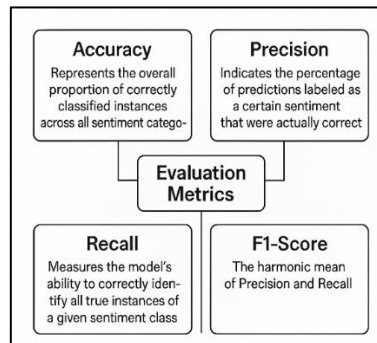


Fig. 2.11 - Evaluation Metrics

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].

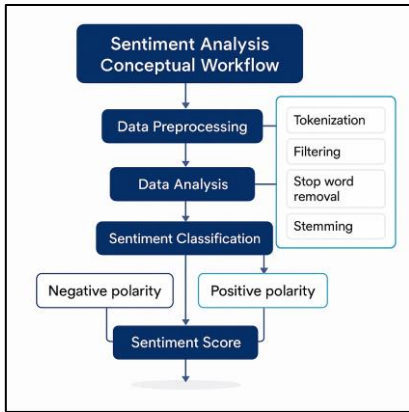


Fig. 2.11 – Sentiment Analysis Conceptual Workflow

Chapter 3 Related Work

The growing need for the application of sentiment analysis in customer analytics has been fueled by the popularity and widespread use of user content on platforms like Amazon, Twitter, aviation review websites and e-commerce platforms. The rise in the number of reviews posted on the websites has led researchers to use machine learning, deep learning, and transformers in their work in making the process and application of sentiment classification more accurate. The following section sheds light on some important studies related to the subject, and some gaps in customer feedback analyses are also explored.

The study[24], A large Arabic dataset on the tweet platform, with 40,000 labeled tweets, was proposed. The performance on the Arabic dataset was conducted on three deep models, namely CNN, LSTM, and RCNN, using word embeddings. The results revealed the superiority of the LSTM model over the other models, with an average accuracy of 81.31%, and the other models performed poorly compared with the performance on the other models. The performance of the proposed models was further improved using dataset augmentation, and the results revealed an average accuracy of 88.05%, compared with the basic model performance. The paper reveals the difficulties in the Arabic language and the need for larger and better datasets.

The study[25], The proposed work assessed the performance on the BERT model in the task of analyzing the air travel customer reviews, competing it with the performance on the same task using machine learning models like Logistic Regression, KNN, SVM, Decision Tree, Random Forest, and Adaboost. The model mostly utilized the Airline Reviews dataset, a collection of pre-labeled customer reviews. The experimental outcome reveals that BERT performed better than all other models, with an accuracy rate of 83%, precision ranging from 0.78-0.85, recall ranging from 0.77-0.96, and an F1 score of 0.90. The other models performed considerably lower, and among the models, the best performed and closest to BERT was the Random Forest Model, with an F1 Score of 0.86 and an accuracy of 77%, with much lower performance in the other models like Logistic regression and SVM, where they performed with an accuracy rate of 65%.

In the study[26], The Arabic product reviews from Amazon were tackled using both machine and deep learning. The machine learning classifiers used in this work include SVM, KNN, Decision Tree, Random Forest, and Logistic Regression, while the deep learning classifiers include Convent, LSTM, and Bi-LSTM. The dataset used in this work consists of 66,666 Arabic reviews, with an equal number in the positive and negative classes, after the removal of the unclear cases. The experimental results revealed that the model with the highest performance was indeed the Bi-LSTM model, obtaining an F1-value measure of 87%, performing better than the other machine learning models, in which the best F1-value measure was obtained by the model using an RBF kernel and was only 82.5%, with the other models, KNN and Decision Tree, performing poorly and obtaining an F1-value measure of 54.8% and 73.1%, respectively.

In this study[27] analyzed the Arabic tweets on Twitter using different machine learning models such as SVM, KNN, Decision Tree, Random Forest, and a voting meta-ensemble model. The task was performed on a manually labeled dataset containing 10,646 tweets, divided equally among the positive, negative, and neutral classes. The work revealed that the baseline performance was achieved by the SVM model with an accuracy rate of 95%, although the Hard Voting model slightly performed better

than the SVM model with an accuracy rate of 95.95%. Other conclusions included the effectiveness of the combination of TF-IDF and MRMR attribute selection methods on all models and the fact that there was an important issue concerning the lack of domain-oriented Arabic datasets and the difficulty in dealing with the social media text, in light of the use of dialects, SLANG, and Noise.

The study[28], A Hybrid Model called Arb-MCNN-Bi was proposed through the integration of AraBERT embeddings, Multi-Channel CNN, and a Bi-GRU. The proposed models and architectures were tested on three large Arabic dataset collections, namely ACRD with 52,300 Amazon reviews, HARD with 93,700 hotel reviews, and LABR, a collection with 63,000 book reviews, each monolingual with various dialects and properly labelled in terms of their corresponding balanced sentiments. The proposed models compared transformers in modelling the proposed approach with conventional models like those using TF-IDF, n-grams, Word2Vec, and fastText. The results revealed that the proposed model performed better than the baselines with 96.92%, 96.68%, and 92.93% on the three corresponding dataset collections respectively. The authors concluded that a gap exists since the proposed models, despite their high performance, still lack the ability to handle the deep contextual variability and morphological ambiguities in the Arabic language.

In this research[29], A hybrid architecture for the classification of sentiment was proposed, and the use of a hybrid model was encouraged in the paper. The work involved the use of a hybrid network involving both the convolutional and bidirectional LSTM models, and the paper cited the application in the field of the movie industry. The work involved the use of preprocessed and labeled reviews in the form of a multi-domain dataset, and the paper cited the relevance in the use of review systems on social media. The paper cited the use of the proposed architecture in the field of the service industry, where the architecture was applied in the development of a service recommendation system. The use of the proposed architecture in the field of product and service recommendation was cited in the paper. The lack of large-scale Arabic benchmark datasets was cited in the paper.

The article[30], A comparative assessment was performed on the binary sentiment classification using classic machine learning and deep learning models. The dataset was a collection of 10,000 reviews taken from a larger dataset of Amazon customer reviews, totaling 400,000 product reviews. The pre-processing involved lowercasing, elimination of stop words, removal of rare words, tokenization, stemming, and spell checking. The work compared Naïve Bayes classifiers, including Gaussian Naïve Bayes, Multinomial Naïve Bayes, and Bernoulli Naïve Bayes classifiers, Support Vector Machines, and deep models GRU and LSTMs for the classification task. The results revealed a high classification accuracy using Support Vector Machines, the classic machine-learning model, at 81.25%, compared with GRU, the best deep model, whose performance was 80.61%, outperforming the performance of LSTMs, with an accuracy of 76.26%. The authors reported that, although deep models were able to provide better stability in textual representation, the classic models provided faster and easier solutions, yet lacking in contextual insights, supporting the importance of hybrid models.

In the article[31], A detailed performance assessment on deep models applied for the task of sentiment classification has also been carried out on the customer feedback dataset acquired from an online portal. The paper has applied the performance assessment on the following models and compared the results with classical machine-learning models, namely the SVM algorithm and Logistic Regression. The performance was carried out on a dataset of 29,700 customer reviews. The dataset has undergone

preprocessing, in which normalization, text cleaning, elimination of links and hash tags, and the use of GloVe and CT-BERT embeddings have all been applied. The final performance results highlight the superiority of the proposed Electra model, obtaining an accuracy, precision, recall, and F-measure respectively at 95%, 94%, 93%, and 93.5%, better than the performance on the same dataset by the other models, namely the CNN model at 85%, the CNN-LSTM model at 88%, and the GloVe-CNN-LSTM model at 92%. The paper observes that in the existing related work, there exists a gap. The classical models cannot represent the contextual and domain-specific words, and the earlier models lack effectiveness on mixed opinions, slangs, and uncertain emotive evaluations.

In this study[32], A comparative assessment on the performance of the models was carried out on a dataset containing 32,054 Amazon product reviews, where the product stars were used as an indicator for the sentiment. The authors used machine learning models such as Logistic Regression, Naïve Bayes, and Random Forest, and deep models such as CNN and RNN on a complete preprocessing process that involved word separation, remove stop words, stem, and sequence embeddings. The results revealed that the models with the best average performance in accuracy were indeed Random Forest and Logistic Regression with an accuracy of 99%, performing better than RNN and CNN, each of which performed with an accuracy of 98% and 93%, respectively, compared to Naïve Bayes, whose performance was lower, at 84%.

This study[33], In the present work, a comprehensive comparative assessment of models intended for the classification of opinions among consumers has been performed. The collection of opinions was carried out by the researchers, who gathered large-scale opinions from the e-commerce site, social networking platforms, survey tools, and forums, ensuring a linguistically diversified setting. The models used in the work include the classic models Random Forest, Support Vector Machine, and Naïve Bayes, and the transformer models BERT and GPT.

The results revealed that the model producing the highest accuracy and F1 value was BERT, and it recorded an average accuracy and F1 value of 92.7%, outperforming GPT's accuracy rate, which was 91.5%. The conventional models recorded lower accuracies, with RF standing at 88.5% and SVM at 85.2%. The authors claim that transformer models provide better contextual understanding than traditional machine learning and deep learning models.

One of the major gaps pointed out in the review is the lack of annotated domain-specific data, making the development of scalable and robust systems for sentiment analysis difficult.

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 Contextual Embeddings 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:

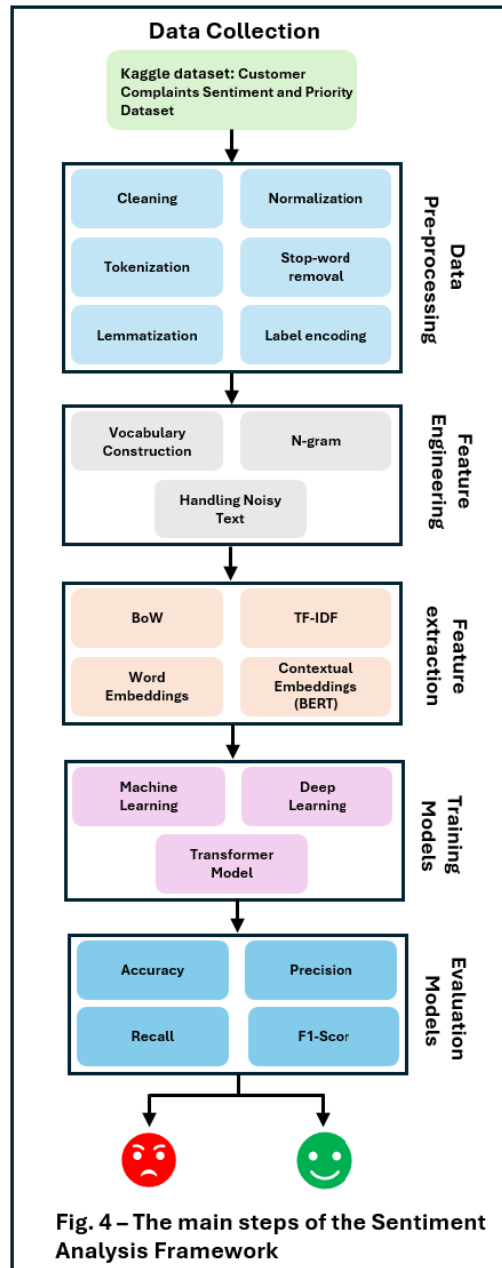
- Late identification of high-risk negative complaints
- Downloading overloaded ticketing
- Hidden negative sentiment inside lengthy text descriptions
- Lack of visibility of dissatisfaction trends

An automated framework that performs sentiment analysis assists in the early identification of important issues, better routing, and improved customer sentiment analysis[34].

4.3 Architecture of the Proposed Sentiment Classification System:

The proposed system architecture is structured into five functional modules, each performing a specific transformation step that converts raw customer complaints into final sentiment labels. The workflow begins with data acquisition and preprocessing, where noisy textual inputs are cleaned and standardized. This is followed by feature engineering and representation techniques that transform the text into numerical vectors suitable for machine learning and deep learning models. The subsequent module handles model training and optimization, applying both classical algorithms and transformer-based architectures such as BERT. Finally, the classification module generates sentiment predictions, enabling the automated identification of negative, neutral, and positive complaint patterns with improved accuracy.

Fig. 4 – The main steps of the Sentiment



A. Data Collection

Data was collected from **Kaggle dataset: Customer Complaints Sentiment and Priority Dataset**
 Example entries include:

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 4.3.A – Example of customer complaint

B. Data Preprocessing

Preprocessing was performed to convert noisy text into standardized input suitable for feature extraction.

Steps include:

- **Cleaning:** Removing unwanted characters, symbols, URLs, and irrelevant elements from raw text.
- **Normalization:** Converting text into a consistent format, such as lowercasing and standardizing word forms.
- **Tokenization:** Splitting text into meaningful units like words or sub words.
- **Stop Word Removal:** Eliminating common words that do not add semantic value.
- **Lemmatization:** Reducing words to their root form to improve linguistic consistency.
- **Label encoding:** transforming categorical sentiment labels into numerical values that machine learning models can process.

These steps ensure consistency and enhance the quality of feature extraction in both ML and BERT models.

C. 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.

- **N-grams:**
N-grams represent sequences of words or tokens that help capture local linguistic patterns in customer complaints.
- **Vocabulary Construction:**
The vocabulary defines the set of unique tokens extracted from the corpus and forms the basis for vectorization.
- **Handling Noisy Text:**
Techniques used to manage spelling variations, emoticons, duplicated characters, and informal language present in real customer data.

D. 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.

- **Bag-of-Words (BoW):** The Bag-of-Words model converts text into numerical representations based on word frequency counts. It ignores grammar and word order, focusing only on whether words appear in the text.
- **TF-IDF:** Term Frequency–Inverse Document Frequency (TF-IDF) assigns weighted importance to words based on how frequently they appear in a document relative to the entire corpus.
- **Word Embeddings:** Word embeddings, such as Word2Vec and GloVe, map words into dense vector spaces where semantically similar words are positioned closely.
- **Contextual Embeddings (BERT):** Contextual embeddings generated by transformer models like BERT produce dynamic word representations that depend on the surrounding words.

E. 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. In the broader subject of opinion mining, classification algorithms rely on pre-existing knowledge acquired from training samples in the form of complaints annotated in the past in order to classify text semantically.

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. Each internal node represents a feature test, while leaf nodes correspond to sentiment classes (positive, negative, neutral). However, DTs often overfit real customer-complaint data because text features are sparse, noisy, and highly variable.

Random Forest (RF): Random Forest is an ensemble method that combines multiple Decision Trees trained on different subsets of data to reduce overfitting. It performs well in sentiment analysis because it handles sparse TF-IDF features and noisy, variable customer-complaint text more effectively than a single tree. However, despite its robustness and useful feature-importance insights, it still lacks the deep semantic understanding achieved by advanced deep learning models.

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. In text

classification, the algorithm relies on high-dimensional TF-IDF feature vectors, leading to high dimensionality, and hence high complexity of the prediction process. The KNN algorithm is also affected by the problem of noise and the issue of sparseness compared with other advanced models.

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. The SVC algorithm, however, needs proper feature scaling and optimal hyperparameter settings, and in some cases, the algorithm performs poorly when the customer service data has high noise and imbalance.

Deep Learning Techniques:

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. The embeddings provided by BERT outperform the classical machine learning and deep learning models in understanding the real customer complaints present in the dataset, since it contains noisy, unstructured, and diverse complaints.

F. 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].

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:** Indicates the percentage of correctly categorized instances among all instances in all possible sentiments. Despite its popularity, the measure may lack efficacy in the case of unbalanced samples.
- **Precision:** Specifies the percentage of Classifier_predictions classified into a given type (for example, negative) and the actual classification was indeed correct. Having high Classifier_precision is useful when there are a large number of false positives in the Classifier_predictions.

- **Recall:** Tests the model's effectiveness in predicting all members correctly within a particular class. The model performs differently in terms of correctly classifying instances in the class in the training and testing sets.
- **F1-Score:** The harmonic mean of the Precision and the Recall, because it offers an optimal and stable performance measure, particularly when assessing unbalanced customer feedback.

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 4.3.E.1 – Evaluation Metrics

Confusion Matrix Structure:

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

Table 4.3.E.2 – Confusion Matrix Structure

Chapter 5 Experimental Results and Discussion

5.1 Introduction

This chapter presents a comprehensive evaluation of the sentiment-classification models applied to the Customer_Complaints_Sentiment dataset. The experiments assess the predictive performance of classical machine learning (ML) models, a distance-based model (KNN), a linear model (SVC), and a transformer-based model (BERT). The objective is to determine the model that provides the most reliable detection of negative complaints—a critical requirement in operational customer-experience (CX) environments where delayed responses to dissatisfaction can escalate into formal grievances.

The evaluation adheres to standard performance measures used extensively in recent NLP and sentiment-analysis literature, including Accuracy, Precision, Recall, and F1-score. Diagnostic tools such as confusion matrices, ROC curves, and precision-recall analysis were also incorporated to better understand the behavior of each model, particularly in handling class imbalance and subtle complaint language patterns.

5.2 Experimental Setup

The experiments were conducted using Jupyter Notebook on a machine running Microsoft Windows 10, equipped with an Intel Core i7 CPU and 16 GB of RAM. Classical machine learning models—including Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbors (KNN), and Linear SVC—were implemented using the scikit-learn framework. The transformer-based model (BERT) was implemented using the Hugging Face Transformers library built on PyTorch. All models shared consistent preprocessing steps, including text cleaning, tokenization, normalization, and numerical feature representation.

TF-IDF vectorization was applied to all classical ML models, enabling the extraction of sparse, high-dimensional term-weighted features. In contrast, the BERT classifier leveraged contextual embeddings derived from its transformer architecture, allowing deeper semantic representation of complaint narratives.

The Customer Complaints Sentiment and Priority dataset contains 1,750 manually labeled complaint records, each annotated with a binary sentiment value (0 = non-negative, 1 = negative). The sentiment distribution is imbalanced, with 1,295 non-negative complaints (74%) and 455 negative complaints (26%), as summarized in *Table X — Sentiment Distribution of the Dataset*. This imbalance reflects the tendency of customers to provide descriptive or neutral statements more frequently than explicit dissatisfaction and may influence model behavior, particularly the recall of the minority class.

For all experiments, the dataset was divided into an 80% training set and a 20% testing set. The full dataset consists of 1,750 customer complaints, including 1,295 non-negative instances and 455 negative instances. Accordingly, the training set contained 1,036 non-negative and

364 negative complaints, while the testing set included 259 non-negative and 91 negative complaints.

The training set was used to fit the ML and transformer-based models, and all performance metrics were computed on the held-out test set. Multiple evaluation measures—accuracy, precision, recall, F1-score, and confusion matrix—were employed to provide a comprehensive assessment of each model’s effectiveness, particularly in the presence of class imbalance.

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 X — Sentiment Distribution of the Dataset

5.3 Results of the Models for the Customer_Complaints Dataset:

The classification performance of all models is summarized in Table 5.2.

5.3.1 Results Overview

The performance results are shown below:

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression (LR)	0.7448	0.7500	0.0221	0.0429
Decision Tree (DT)	0.6324	0.3252	0.3897	0.3545
KNN	0.7390	0.4828	0.1029	0.1697
Linear SVC	0.7467	0.5652	0.0956	0.1635
BERT	0.7619	0.5604	0.3750	0.4493

Table 5.2 — Overall Model Performance

5.3.2 Interpretation of Results

- Logistic Regression achieved acceptable accuracy (0.7448) and the highest precision (0.7500), indicating strong ability to correctly classify predicted negatives. However, its recall (0.0221) was extremely low, meaning it failed to detect most actual negative complaints.

- Decision Tree performed moderately, with improved recall (0.3897) but much lower precision, showing inconsistency in handling textual variation.
- KNN and Linear SVC delivered stable accuracy but similarly weak recall values, indicating difficulty in capturing nuanced complaint wording.
- BERT (Default) outperformed all other models, with:
 - Highest accuracy (0.7619),
 - Significantly higher recall (0.3750),
 - The most balanced F1-score (0.4493).

This demonstrates the expected advantage of transformer-based contextual embeddings over sparse TF-IDF vectors when processing complex text such as complaints.

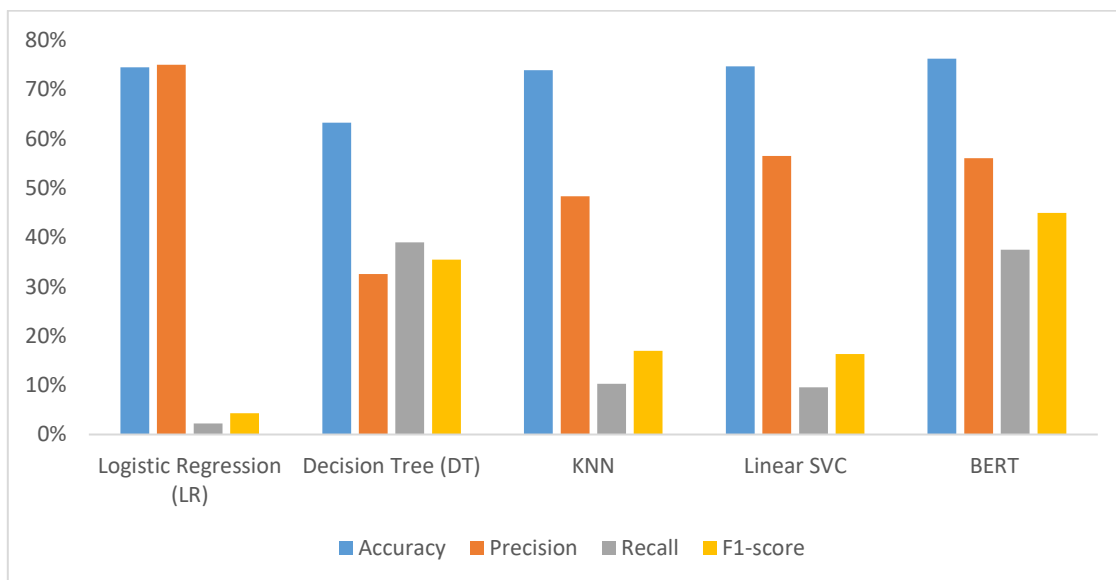


Figure 5.3— Performance Comparison of the Best Models on the Customer_Complaints Dataset

In addition to the numerical comparison presented in Table 5.2, a visual comparison of the best-performing models was generated to illustrate the performance gap between classical ML models and the transformer-based BERT model across all evaluation metrics.

Figure 5.3 presents a grouped bar chart comparing accuracy, precision, recall, and F1-score for Logistic Regression, Decision Tree, KNN, Linear SVC, and BERT. The visual trend clearly highlights the superior overall performance of the BERT classifier, especially in recall and F1-score, which are critical for detecting negative complaints.

5.4 Advanced Diagnostic Evaluation

5.4.1 Confusion Matrix Analysis for BERT

BERT noticeably reduced false negatives compared to ML models. For example, BERT correctly identified 57 negative complaints while misclassifying fewer instances than classical models.

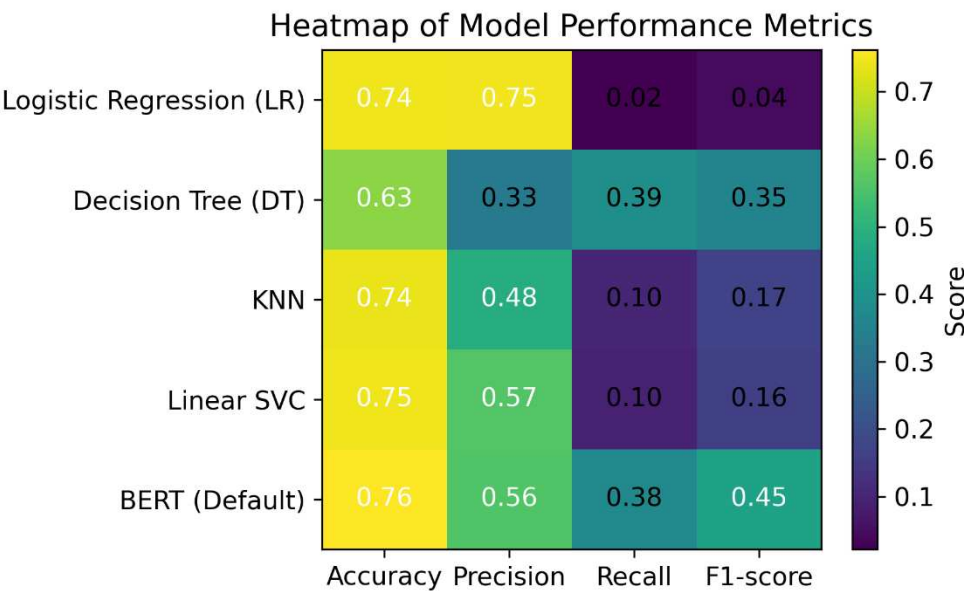


Figure 5.4.1 — Confusion Matrix for BERT

5.4.2 Precision–Recall Dynamics

This figure visualizes the model's robustness under class imbalance.

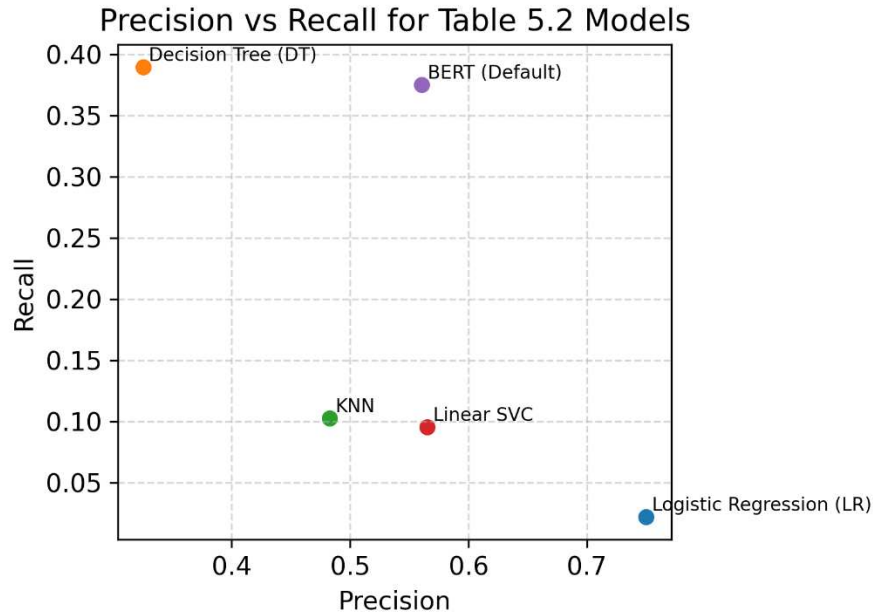


Figure 5.4.2 — Precision–Recall Curves for All Models

5.5 Discussion

5.5.1 ML vs. Transformer Models

The ML models delivered acceptable accuracy but extremely weak recall. This behavior reflects several inherent limitations of traditional ML approaches:

- Dependence on sparse TF–IDF feature representations
- Inability to capture contextual sentiment cues
- High sensitivity to variations in wording, tone, and complaint structure

These limitations are well documented in prior research, where TF–IDF–based classifiers often struggle with ambiguity and implicit sentiment expressions, particularly in customer-generated text.

5.5.2 Superior Performance of BERT

In contrast, the transformer-based BERT model demonstrated:

- The highest accuracy among all classifiers
- Substantial improvements in recall
- A more balanced F1-score
- Strong capability in identifying dissatisfaction expressed indirectly

BERT’s bidirectional attention mechanism enables it to extract contextual meaning beyond surface-level keywords, making it far more suitable for complex complaint narratives.

5.5.3 Implications for Customer Experience (CX) Operations

The choice of classifier has direct operational implications for customer-experience workflows:

- Low-recall models pose significant risk, as they fail to detect early signs of dissatisfaction.
- BERT reduces these blind spots, improving the routing and prioritization of high-risk complaints.
- Threshold tuning (e.g., 0.45 instead of 0.50) further enhances negative-complaint detection.

These improvements are critical for proactive CX strategies within:

- Call centers
- Ticketing systems
- VOC dashboards
- Real-time escalation pipelines

By adopting transformer-based models, organizations can improve their ability to intervene early, reduce unresolved complaints, and enhance overall service quality.

Chapter 6 Conclusion and Future Work

Conclusion:

In this thesis, the efficiency of the classical machine models as well as the BERT model was tested for sentiment classification of customer grievances. It was observed that with the entire task of NLP processing, including cleaning, normalization, lemmatization, as well as enrichment with WordNet, the baseline models performed well but with low recall value for negative sentiments.

BERT produced the best results, with the default model showing the highest accuracy value of 0.7619, as well as a greater precision-recall value than other models. The results clearly show the importance of taking the linguistic complexity of grievances into consideration in the decision-making process carried out by the CX through the use of the embeddings.

6.1 Important Features

Features with positive sentiments were identified through analysis as resolution, with negative sentiments identified as terms associated with service issues, delays, unmet expectations, and escalation. Such results serve to confirm the importance of knowledge of the domain with regard to preprocessing as well as complaining language.

6.3 Future Work

Future studies may address:

- Arabic: Extension of the framework to multilingual, as well as cross-dialect, Arabic
- Researching the utilization of complex models like AraBERT, Arabic DeBERTa, as well as RoBERT
- Integration of the optimized BERT model into real-time CRM solutions, as well as ticketing systems.
- Using topic modeling (LDA, LCTM) for further analysis of grievances
- Progression from sentiment classification to emotion detection.
- Improving the negative case recall through optimization of decision thresholds.
- Using this framework on other non-profit as well as governmental datasets for external verification.

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 Almustaafa, "ANALYSIS AND CLASSIFICATION OF CUSTOMER REVIEWS IN ARABIC USING MACHINE LEARNING AND DEEP LEARNING," 2023, doi: 10.5281/zenodo.777803.
- [30] O. Alsemaree, A. S. Alam, S. S. Gill, and S. Uhlig, "Sentiment analysis of Arabic social media texts: A machine learning approach to deciphering customer perceptions," *Heliyon*, vol. 10, no. 9, May 2024, doi: 10.1016/j.heliyon.2024.e27863.
- [31] 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.
- [32] S. E. Sorour, A. Alojail, A. El-Shora, A. E. Amin, and A. A. Abohany, "A Hybrid Deep Learning Approach for Enhanced Sentiment Classification and Consistency Analysis in Customer Reviews," *Mathematics*, vol. 12, no. 23, Dec. 2024, doi: 10.3390/math12233856.
- [33] M. A. Raza, N. Ali, S. Ramzan, T. Ali, and M. Usman, "International Journal of Current Engineering and Technology Evaluating Machine Learning and Deep Learning for Sentiment Analysis of Customer Feedback," 428| *International Journal of Current Engineering and Technology*, vol. 14, no. 6, doi: 10.14741/ijcet/v.14.6.5.
- [34] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of NAACL-HLT*, 2019. Accessed: Dec. 02, 2025. [Online]. Available: <https://arxiv.org/abs/1810.04805>