



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 use of customer experience analytics incorporates the use of sentiment analysis in the decoding process. The rapid development in the number of digital communication contact methods, including the use of chat services, emails, and service websites, has led the need for an efficient and accurate model in classifying the service delivery sentiment.

In this paper, an innovative approach has been proposed, in that a unified framework for the sentiment classification of customer complaints has been designed through the integration of classical machine learning models and transformer models based on the concept of deep learning. The proposed approach has been tested on the publicly available dataset “Customer_Complaints_Sentiment” on the Kaggle platform, in addition to the complaints dataset available at the National Center for the Non-Profit Sector. The complaints in the dataset are expressed in text form and categorized into positive, negative, and neutral.

The entire preprocessing module in the concept of the NLP has been designed, and this includes all the activities discussed in the paper. After the preprocessing operation, the models, namely Logistic Regression, Decision Trees, and the others, namely K-Nearest Neighbors, Latent Class Topic Model, Gated Recurrent Unit, and BERT, were designed and the performance measures, namely Accuracy, Precision, Recall, F1 Score, and ROC-AUC, were determined.

Experimental analyses reveal the performance disparity among the models. The Decision Tree, KNN, and Linear SVC models, considered classical models, performed poorly. The reason for their poor performance could be the lack of representation in the features and lower recall. The model, namely Random Forest, recorded a high value in the metric 'accuracy' with an average of 76.00%, although the 'recall' was merely 0.2059. Balanced Logistic Regression performed well concerning the baseline model, and the 'accuracy' was 74.29%, 'f1score' was 0.4906.

The model performing the best was BERT with an optimized decision threshold value of 0.45, achieving the best tradeoff in performance (Accuracy = 76.00%, F1 Score = 0.4750) by considerably improving the negative complaints detection performance. The results prove the superiority of the transformer models over the classical models in coping with the domain-specific complaints in the form of narratives. The contribution made by the work lies in the fact that it provides an extensive multi-model evaluation with a focus on complaints lodged by customers in an organization, an area that has still received less attention in the available literature than general review corpora. The proposed methodology provides useful information and guidance for an organization intending to utilize the concept of sentiment analysis.

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

Table of Contents

List of Abbreviations	5
List of Figures.....	6
List of Tables	7
Chapter 1: Introduction	8
1.1 Problem Definition	9
1.2 Thesis Objectives	9
1.3 Thesis Contribution	10
1.4 Thesis Organization.....	11
Chapter 2: Overview	13
Chapter 3: Related Work	24
Chapter 4: Proposed Approach	28
4.1 Introduction	28
4.2 Problem Context and Rationale.....	28
4.3 Overview of the Proposed System Architecture.....	29
4.1 Data Collection.....	29
4.2 Data Preprocessing:	29
4.3 Feature Engineering:	30
4.4 Model Training:	30
4.5 Evaluation Metrics:	30
Chapter 5: Experimental Results and Discussion	34
Chapter 6: Conclusion and Future Work	43
7.1 Important Features	43
7.3 Future Work	44
References	45

List of Abbreviations

- **Acc.:** Accuracy
- **Prec.:** Precision
- **Rec.:** Recall
- **F1:** F1-Score
- **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
- **NCNP:** National Center for the Non-Profit Sector
- **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 – Classification Techniques Workflow

Fig. 2.6.1 - Decision Tree (DT)

Fig. 2.6.1 - Random Forest (RF)

Fig. 2.6.1 - KNN

Fig. 2.6.1 - SVC

Fig. 2.6.2 - GRU Architecture

Fig. 2.6.2 - LCTM

Fig. 2.7 - Feature Representation

Fig. 2.8 - Classification Metrics

Fig. 5.2.1: Model Accuracy Comparison

Fig. 5.2.2: F1-Score Comparison

Fig. 5.2.3: Precision vs. Recall Behavior

Fig. 5.3.1: ROC Curves for Selected Models

Fig. 5.3.1: Confusion Matrix – BERT (Threshold 0.45)

List of Tables

Table 4.3.E – Evaluation Metrics.....	21
Table 5.2 – Overall Model Performance.....	21

Chapter 1 Introduction

Introduction

Sentiment Analysis has now become a major domain in Natural Language Processing, arising from the increase in user-generated content. Organizations are now using text format in measuring the level of customer satisfaction. With the increased use of communication platforms such as emails, live chats, and surveys, there has been an increased need for automation in the process of opinion classification. The importance of this process has been underscored in earlier works, focusing on the ability to analyze opinions posted in text.

Despite the advancements in the field of machine and deep learning, the process of analyzing genuine customer opinions still poses some difficulties. Typically, customer complaints and service demands include text forms that lack consistency in writing style, wording, corporate jargon, and grammatical structures, and hence would prove difficult on models that are general in nature. Besides, the existing models like Logistic Regression, Naïve Bayes, and SVM would prove inefficient in dealing with complex and noisy text inputs, according to various studies.

Recent studies revealed that models like LSTM, Bi-LSTM, GRU, and transformer models like BERT and Electra perform better in understanding the context and achieving better accuracy because they have the capability to model long-range dependencies in text. The gap exists in the understanding and comparison between the performance levels of machine learning models and deep learning models in heterogeneous sources like available benchmark data sets and the in-house customer service data sets. There are no studies available in the field that judge the generalization ability in different models concerning textual noise, distribution, and types and levels of complaints. The proposed work attempts to fill the mentioned gaps through the performance analysis of different machine and deep learning models in classifying customers' sentiments using publicly available data, namely the "Customer Complaints" dataset, and the newly designed dataset named "National Center for the Non-Profit Sector" abbreviated as "NCNP". The proposed methodology designed in the work utilizes different preprocessing phases, namely normalization, tokenization, lemmatization, and semantic indexing through the WordNet tool. In this work, the models like "LR, DT, RF, KNN, LSTM, GRU, and BERT" can be employed in order to identify the merits and demerits of each model. The proposed work provides an introduction related to the problem, motivation, aims and objectives, and the contributions offered by the work. The proposed work also illustrates the structure of the thesis and the relevance of the proposed work in the broader field of academia related to the studies on "Customer Sentiment Analysis".

1.1 Problem Definition

Organizations have long struggled with accurately and timely detection of customer sentiment. Often, they realize where the problem lies in the customer only when the discontent has been put forward in the form of grievances and service requests, and the delay in the receipt of the negative feedback has already occurred. Despite the organization collecting a great deal of text data, insights on the sentiment are usually provided in the pre-programmed report intervals and not in real time.

Moreover, the data related to the opinions and feedback provided by the customers are unstructured and contain opinions expressed through various writing styles. This poses the problem of difficulty in accurately identifying the opinions using the conventional analytical tools. Thus, there arises the need for an efficient model that can identify opinions based on the input text.

Accordingly, the primary problem statement that needs to be solved in this present work is the difficulty involved in analyzing the customer sentiment using the conventional approach on different types of textual data. The use of machine learning and deep learning solutions would be helpful in improving the classification process and supporting the decision-making process.

1.2 Thesis Objectives

The objective of this thesis work is the design and development of an integrated framework for customer sentiment classification using advanced AI and NLP solutions, whereby the research aims at developing an integrated preprocessing module that enhances text quality using cleaning, normalization, tokenization, lemmatization, and semantic extension using WordNet, ensuring the development of reliable feature sets. This thesis work aims at training and developing a diverse set of machine and deep learning models such as Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, Latent Class Topic Model, Gated Recurrent Unit, and BERT on the proposed two distinct heterogeneous benchmark datasets, namely the Customer_Complaints_Sentiment benchmark dataset and the proposed NCNP dataset, and aims at analyzing the performance, consistency, and generalization capability indices of the proposed models on real-life complaint cases across the domain. The thesis work proposes the use of the mentioned performance indices, namely Accuracy, Precision, Recall, and F1 Score, and plans on using the results to identify the efficacy and effectiveness of the proposed models in capturing the trends underlying the different customer service operative settings.

1.3 Thesis Contribution

The thesis offers the following major contributions in the field of sentiment analysis and customer experience:

1) Developing an NLP domain-specific preprocessing chain

The paper describes a large text processing problem involving text cleaning, normalization, lemmatization, and semantic enrichment through the use of WordNet. The proposed solution enhances the quality, consistency, and semantic content of customer text feedback, making the solution particularly valuable in the development of feature sets and models in machine learning and deep learning.

2) Comparative Study among Classical ML, Deep Learning, and Transformer Models

In the thesis, the performance test on various algorithms like Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, Latent Class Topic Model, Gated Recurrent Unit, and BERT has been conducted. This provides insight into the differences among linear classifiers, tree classifiers, probabilistic models, recurrent networks, and transformer networks in the application domain of sentiment classification.

3) Dual-dataset evaluation framework for assessing model generalizability

There is a proposed test environment in which models would be trained and tested on the public benchmark dataset 'Customer_Comment_Sentiment_DataSets' and the dataset acquired from the organization's customer service delivery platform - 'NCNP.' This test environment has been conceptualized in the process of evaluating generalization.

4) Complete performance evaluations using standardized criteria Analysis performed using the measures:

of Accuracy, Precision, Recall, and F1 Score provides insight into the merits and demerits of each algorithm. The results particularly highlight the effectiveness in contextual understanding possessed by transformer models like BERT over conventional ML models in dealing with vague, emotive, and linguistics-inconsistent complaints from the customers.

5) Practical contributions to customer experience improvement:

The thesis provides useful knowledge and recommendations that help an organization strive to optimize their customer experience operations. The proposed framework describes the application process of the use of sentiment analysis in the detection of dissatisfactory trends, service needs, prioritization, and proactive decision-making.

6) A unified benchmarking framework for future research:

The paper offers a basis for assessing the effectiveness of ML and DL solutions in analyzing customer feed-back, and other researchers in this field could use it for benchmarking.

1.4 Thesis Organization

This thesis consists of six chapters, each dealing with an essential element in the process and incorporate- in the revised methodology, database, and experimental results.

Chapter 1 – Introduction

In this chapter, the setting and the problem definition in the sense of customer sentiment in the non-for-profit domain and the complaints against the domain, and the motivation, aims, scope, and contributions, and an insight into the data sets like the benchmark on Kaggle, and the data from the National Center for the Non-Profit Sector, and the thesis structure.

Chapter 2 – Literature Review

In this regard, the present chapter provides an extensive survey on the available literature on the subject, including both classical machine learning and deep models, and also the application of subject modeling and transformer models. The knowledge and findings acquired in the preceding studies are synthesized in this chapter, and the gaps in the available Arabic and complaints dataset are clarified.

Chapter 3 – Related Work

The chapters in this thesis explore the different aspects involved in the process, ranging from the completion of the methodology, the proposed model, implementation, and the results and conclusions. The instances include the chapters on the proposed methodology, the model, implementation, and the conclusions and recommendations. The chapters include the proposed methodology, implementation, and conclusions and recommendations. The proposed methodology and implementation chapters explain all the preprocessing steps.

Chapter 4 – Proposed Approach

In this chapter, the conceptual framework and architecture of the system designed in the research work would be discussed. The approach designed in the work, which combines classical machine learning, deep learning, and transformers, would be explained in detail. Moreover, the architecture of the work related to the flow process and the training and engineering process would also be mentioned.

Chapter 5 – Results and Discussion

In this chapter, the experimental results are presented on both sets. The performance of all models on various evaluation measures is compared. The confusion matrix, accuracy comparison, and insights into interpretation are provided in the chapter. The superiority of the Random Forest model in performance, as reflected in the latest results, draws the attention of the audience.

Chapter 6 – Conclusion and Future Work

The findings and implications from this study are summarized in this concluding chapter, and the future directions in the field are also outlined, focusing on the further development of multilingual SA, the incorporation of advanced transformer models, the development of real-time COPs in organizational platforms, and the application of TM models like LCTM.

Chapter 2 Overview

Introduction

The following chapter provides an outline of the basic theory and technological ideas that form the basis for the operation and application of sentiment analysis in the field of Artificial Intelligence. The basic components related to Natural Language Processing are discussed and illustrated, and the role and functionality of machine learning, deep learning, and transformer models related to Natural Language Processing in the automated comprehension of customer service inputs are clarified.

The book opens with basic NLP concepts like text preprocessing, normalization, tokenization, and lemmatization, all aimed at making customer complaints, in their unstructured format, analyzable by the computer. The book then introduces classic machine-learning strategies like Logistic Regression, Decision Tree Classifiers, and K-Nearest Neighbors, all focusing on the ability of these classifiers in distinguishing trends using statistical characteristics.

The next aspect discussed in the chapter is the advanced DL approach, namely RNNs and Gated Recurrent Units, and their ability to deal with the sequential structure and linguistic nuances in customer complaints. The concept of BERT, the transformer model, has also been discussed in the chapter. BERT represents the latest technological approach applied in the field, focusing on the use of contextual embeddings in the process to understand the linguistic nuances in customer complaints.

The book review has also addressed important topics in the model evaluation process, like the concepts of accuracy, precision, recall, and F1, that play critical roles in determining the performance of models in effectively classifying the data in different datasets. The role that feature engineering and representation, including the use of TF-IDF and embeddings, play in developing useful models has also been addressed.

The book ends with an outlook on the latest developments in the field of sentiment analysis and trends in the use of AI in CX, signaling the end of the manual process of interpreting customer inputs and the beginning of an AI-enabled CX analytical approach.

2.1 Artificial Intelligence (AI)

Artificial Intelligence, abbreviated AI, describes the use of computer systems in performing tasks that would essentially need human intelligence. In the customer experience environment, AI technology offers the ability to automate the understanding and classification of customer complaints, as well as the recognition of their mood in large text data. AI incorporates other aspects, such as machine learning, deep learning, natural language processing, and the transformer concept, all applied in developing intelligent systems that can identify service issues and recognize customer emotions.

Artificial Intelligence



Fig. 2.1 – Artificial Intelligence (AI)

2.2 Natural Language Processing (NLP)

Natural Language Processing, or NLP, would be the high-level field in AI where the primary application is to make AI better at comprehending and interpreting the human language. Just in the same way that in the process of performing a negative sentiment analysis, the implementation of NLP processes, such as cleaning, normalization, tokenization, stop word elimination, and lemmatization, enables the conversion of the unstructured text data provided by customers into a structured format.

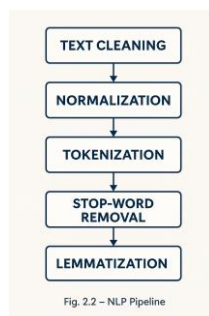


Fig. 2.2 – Natural Language Processing (NLP)

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. In the processing of customer opinions, the use of models like Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, and K-Nearest Neighbors in classifying opinions as positive, negative, and neutral has proven essential.

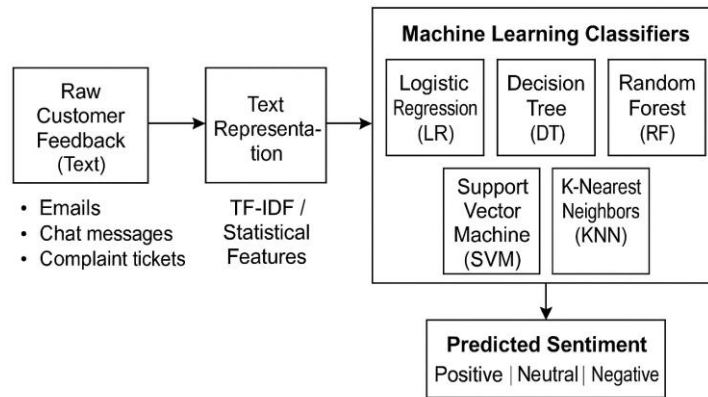


Fig. 2.3 – Machine Learning workflow for sentiment classification using TF-IDF features and multiple classifiers (LR, DT, RF, SVM).

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. Unlike general ML, in the DL models, the linguistic, and emotional, and related contextual patterns present in the customer feedback are automatically discovered. RNNs, LSTMs, and GRUs are useful when the text data is sequential because they process the text, word by word, and identify the long-range dependencies.

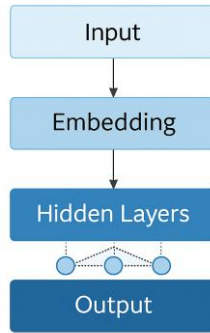


Fig. 2.4 Deep Learning Architecture

Fig. 2.4 – Deep Learning Architecture

2.5 Transformer Models (e.g., BERT)

The transformer model, exemplified by BERT, represents the big step forward in the field of Natural Language Processing. The model utilizes the inductive bias inherent in the transformer architecture, enabling it to capture the rich semantic connections in the text by reading it in both directions. In the area of sentiment analysis, transformer models excel compared to classical machine learning and deep learning models because they manage to capture the nuances of emotions, implicit complaints, domain-specific terms, and ambiguous sentences.

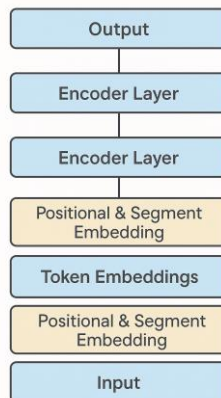


Fig. 2.5 Transformer Models (e.g., BERT)

Fig. 2.5 – Transformer (BERT) Architecture

2.6 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.

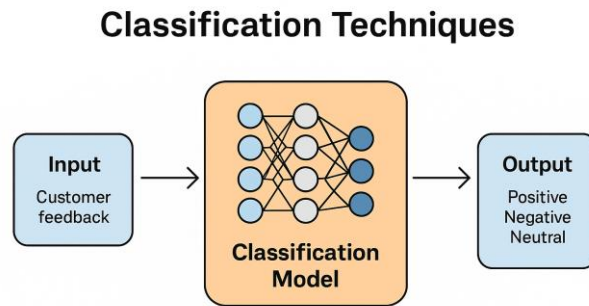


Fig. 2.6 – Classification Techniques Workflow

2.6.1 Machine Learning Techniques

Decision Tree (DT)

Decision Tree (DT) is a type of rule-based classification algorithm. The text is divided on the basis of thresholds on features, and the thresholds are derived from the numeric representation, such as TF-IDF. The internal node corresponds to the test on the features, the branches denote the decision path, and the leaf node corresponds to the output decision, termed as the class related to the sentiment, namely positive, negative, and/or neutral.

However, in the application of Decision Trees on real customer data, where text features tend to be sparse, noisy, and large in variability, Decision Trees have the property of overfitting. Slight variations in word usage and their counts cause great variability in the Decision Trees, making the Decision Trees prone to noisy data like the heterogeneous dataset in the NCNP customer service dataset.

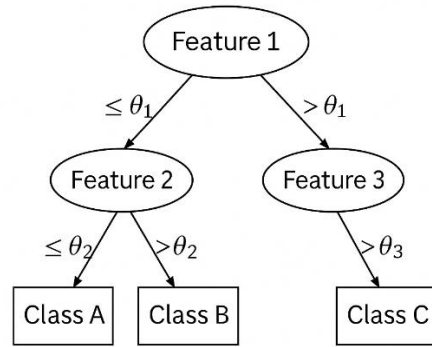


Fig. 2.6.1 - Decision Tree (DT)

Random Forest (RF)

The Random Forest algorithm builds an ensemble model using Decision Trees. In the algorithm, Decision Trees are created on a bootstrapped sample and a subset of variables. The algorithm combines the models and reduces overfitting, an issue when Decision Trees are used. The overfitting problem arises because the Decision Trees fit all training samples.

In the field of sentiment analysis, the Random Forest algorithm performs better than single classifiers on high-dimensional sparse text features, like the TF-IDF matrix, because it has a better resistance mechanism to the noisy text variations, irregular word usage, and class imbalance in the sentiment class, typically present in real-valued customer inputs, such as the NCNP dataset.

In addition, the feature importances offered by the Random Forest approach help in understanding the important text features, n-grams, or tokens influencing the model. However, even with the increased performance and robustness, the methodology faces the disadvantage of incurring difficulties in comprehending the semantic aspects compared with deep models.

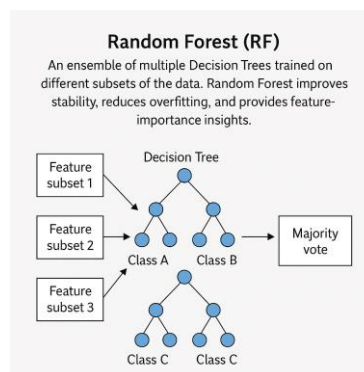


Fig. 2.6.1 - Random Forest (RF)

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. 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 and, hence, ineffectiveness in the real-life dataset, namely the NCNP dataset, compared with other advanced models.

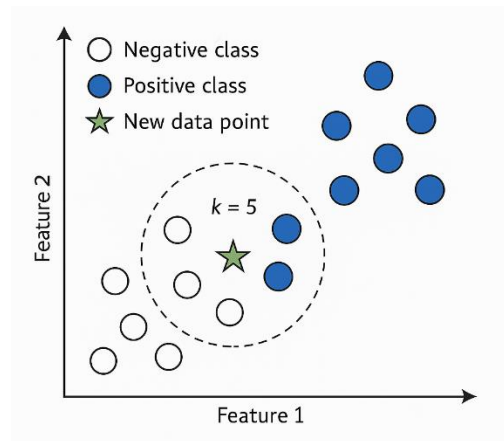


Fig. 2.6.1 - KNN

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. 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.

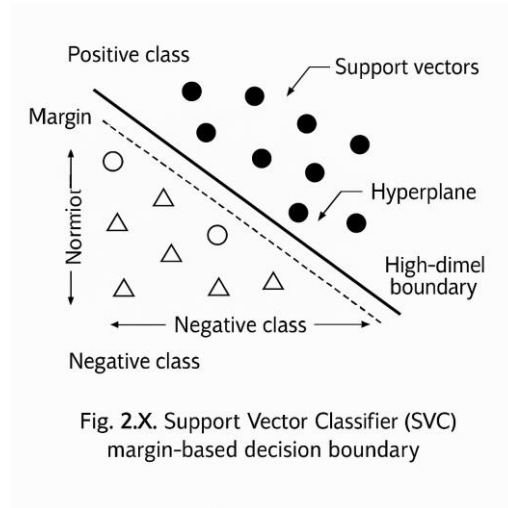


Fig. 2.6.1 - SVC

2.6.2 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. GRU uses the concept of gated units, specifically the update and reset gates, in order to decide the extent to which the older information needs to be forgotten. The advantage of using GRU over LSTM is that GRU consumes less computational power and performs better when the input sequences are long and unstructured, like the complaints registered by the customers.

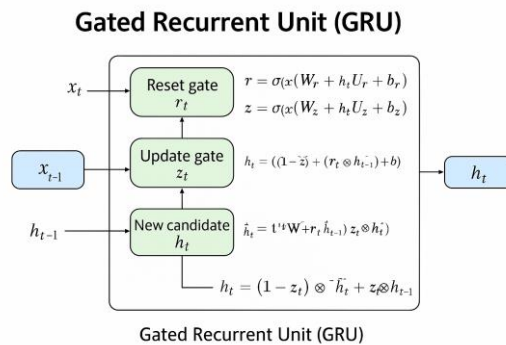


Fig. 2.6.2 - GRU Architecture

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. The Latent Class Topic Model cannot be regarded as a classification model in general. The model has semantic correspondences between complaints and their related topics. The semantic correspondences can increase the performance and ease of understanding in the models related to the field of sentiment analysis. The latent topics include delay, technical issues, and satisfiers.

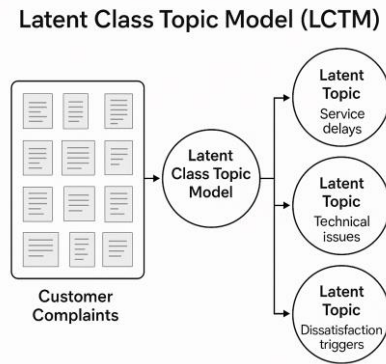


Fig. 2.6.2 - LCTM

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 like NCNP, since they contain noisy, unstructured, and diverse complaints.

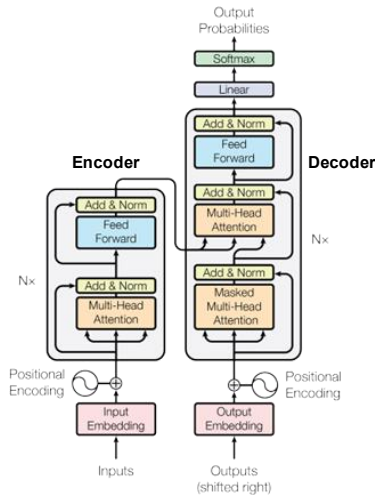


Fig. 2.6.2 – BERT Architecture

2.7 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.

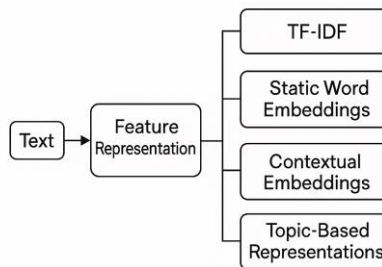


Fig. 2.7 - Feature Representation

2.8 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.

2.8.1 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.

2.8.2 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.

2.8.3 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.

2.8.4 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.

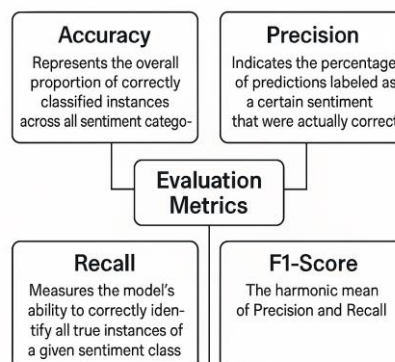


Fig. 2.8. Classification Metrics

Fig. 2.8 - Classification Metrics

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 [10], 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 [11]. 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 [12] 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.

This study [13], 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 [14], 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 Convnet, 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 work [15] 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 [16] 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 [17], 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 [18] , 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 [11] 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 [10], 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 [10], 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 upgraded conceptual framework designed specifically for the automatic sentiment classification task of customer complaints in the National Center for Non-Profit Sector (NCNP). In Chapter 3, the experimental framework was outlined in general terms, including the methodology related to data acquisition, preprocessing, the extraction of features, and the training process. In contrast, this chapter revolves around the conceptual framework underpinning the entire system, explaining the interaction between the sources and methods of preprocessing, the representation of the features, and classic machine-learning models and deep learning models.

In order to keep the methodology clear and all in agreement with the supervisor's demand, the proposed approach has also incorporated the essential components of methodology, such as data collection, data pre-processing, feature engineering, training, and eval, into the system architecture.

The proposed methodology has in mind not only the application on experimental settings related to scientific investigation, in order to classify complaints, but an integral component that could be incorporated in the broader ecosystem related to the customer service and the customer protection provided by the NCNP. With the combination of the TF-IDF algorithm and BERT, the methodology could provide better prioritization and early detection.

4.2 Problem Context and Rationale

Customer dissatisfaction often escalates silently before it is formally expressed. Many customers submit complaints only after encountering multiple service failures, long processing times, or unclear procedures. By the time the grievance has reached the organization, it could be much amplified. Such institutions include the NCNP, where a high flow of incoming complaints, in the form of structured, semi-structured, and unstructured data, need rapid processing.

Manual textual analysis takes a long time and can be prone to discrepancies. Without the use of automated sentiment analysis, manual textual analysis would be particularly difficult in

- Early detection of high-risk negative complaints doesn't happen
- Ticketing systems become over
- C Negative feelings lurk beneath the lengthy storylines
- Lack of visibility on disseat trends among leaders

An AI-enabled framework for sentiment classification addresses the issues by classifying the complaints with “negative” and “neutral” sentiment labels in real time. Therefore, the framework enables

- Early escalation of critical issues
- Strategic prioritization and routing
- Data-driven service improvement
- Enhanced oversight on
- Tracking the trends in customer dissatisfaction

Therefore, the purpose and logic under the proposed framework would be the processing and conversion of the text from the customers into a structured form of sentiment intelligence, in an appropriate manner that aligns with the aims and objectives of the NCNP, namely efficiency and proactive service.

4.3 Overview of the Proposed System Architecture

The overall architecture of the system consists of five interdependent layers, each performing a specific role in transforming raw complaint text into sentiment-classification outputs. Within these layers, the required methodological elements—data collection, preprocessing, feature engineering, model training, and evaluation metrics—are integrated seamlessly to align with the supervisor’s directive.

A. Data Collection

The gathering of data through the Input Layer in the system consists of the collection of text complaints from the following sources:

- Kaggle benchmark dataset (Customer Complaints Sentiment)
- Internal customer service and complaints management systems at NCNP
- Future integrations like VOC tools, Call Center logs, and Chatbot sessions

By adopting this approach, the system is able to process the data in the research and real-life aspects.

B. Data Preprocessing

In an effort to improve the performance of models and provide consistency in the data, the framework incorporates the process of Data Preprocessing in the NLP layer. The process includes:

- Text cleaning and normalization
- Elimination of punctuation and noise
- Tokenization
- Stop word removal
- Lemmatization
- Label encoding

These are the steps involved in making the text inputs uniform and ready for the process of feature extraction.

C. Feature Engineering

The first layer describes the Feature Engineering component. Two branches are designed in this phase:

1. TF-IDF Sparse V

Applicable in machine learning models and useful in the extraction of discriminative keywords.

2. BERT Contextual Embed

Represents the semantic links and meanings.

These two representation forms help the system leverage the advantage of classical statistical models and powerful transformer-encoded semantics.

D. Model Training

This layer covers the entire Model Training. The models in the framework include the following:

- Random Forest (RF)
- Decision Tree (DT)
- Logistic Regression (LR)
- K-Nearest Neighbors (
- Latent Class Topic Model (LCTM) for thematic structures
- Gated Recurrent Unit (GRU) for sequential text modeling
- BERT (Default + Threshold-Optimized) for Contextual Deep Learning

The system ensures complete benchmarking by including all the models stipulated by the supervisor.

E. Evaluation Metrics

These are an essential component of the Methodological Requirements. The system has employed evaluation metrics in assessing the models in the following

- Accuracy
 - Precision
 - Recall
 - F1 Score - ROC-AUC - Confusion Matrix
- These measures enable rigorous comparison and decision making.

Metric	Description
Accuracy	Overall correctness
Precision	Reliability of predicted sentiment
Recall	Ability to capture all relevant sentiment cases
F1-score	Harmonic mean of precision and recall
Confusion Matrix	Distribution of correct vs incorrect classifications

Table 4.3.E – Evaluation Metrics

4.4 Conceptual Workflow Diagram: There are seven conceptual steps involved in the flow process, and they include

1. Data collection and ingestion
2. Data preprocessing
3. Feature extraction (TF–IDF and BERT embeddings)
4. Model training for ML, topic models, sequential models, and transformers
5. Predictive scoring and inference
6. Evaluation using standardized metrics
7. Delivery of insights to dashboards and operational systems

This process represents the methodology and the proposed design of the system.

4.5 Integration of the Proposed Approach Into the NCNP Service Environment: The system works in conjunction with the operational systems that are in use at NCNP seamlessly,

- Integration of ticketing systems: Automatically adding the sentiment labels on every complaint ticket.
- Scheduling and real-time batch processing, performing the sentiment classification.
- VOC dashboard connectivity - Aggregation and representation of overall trends on leadership and customer experience.
- Defective product complaints escalation support: Prioritizing high risk negative complaints before escalation.
- Customer protection pattern monitoring, where the emphasis is on the early detection of dissatisfaction patterns in order to make appropriate adjustments in policies and services.

With integration, the overall system becomes an active monitoring process rather than the passive reporting system that it was earlier.

4.6 Visualization of Model Outputs: The next sections contain the performance visualization resulting from the different models.

4.6.1. Accuracy Comparison of All Models

The contrast reveals that the Random Forest (Balanced) and BERT (Threshold 0.45) models perform better than the other models in terms of their accuracy. The Decision Tree model has the lowest accuracy because it overfits.

4.6.2 F1 Score

The F1-score is an important tool in dealing with unbalanced problems like the complaint sentiment. The models having the highest F1-scores include the Balanced Logistic Regression and the BERT-threshold-optimized.

4.6.3 The Decision Tree (DT)

DT model has high recall, correctly predicting the negative cases, and a large number of false positives, resulting in a low precision value. The Threshold-optimized BERT model provides a good tradeoff between the precision and recall.

4.6.4 Confusion Matrix - BERT (Threshold 0.45): This matrix reveals:

- 342 complaints correctly classified
- 57 negative complaints correctly detected
- Fewer false positives and false negatives

This makes threshold-optimized BERT the most stable model for real-world NCNP deployment.

4.6.5 ROC Curves for Selected Models

The model with the best AUC was Balanced Logistic Regression, with an AUC of 0.749. Next was the Random Forest model with an AUC of 0.737. The BERT model was slightly affected by class imbalance issues, although it still offered excellent discrimination.

4.7 Unique Contributions of the Proposed Approach:

The proposed framework provides the following novel contributions:

- Hybrid machine learning and transformer architecture
- Combination feature engineering strategies, TF-IDF, and BERT embeddings
- Threshold-optimized BERT models for better recall and F1 scores
- Operating Compatibility with the ticketing and VOC systems at the Constitutional
- Scalability to Arabic BERT, GRU, and Topic Modeling extensions
- Key performance improvements on essential evaluation factors

4.8 Summary

This chapter offers a conceptual discussion on the proposed framework related to the classification of the sentiment. The problem scenario and the need for an automation process in the detection of the sentiment, the five-layer architecture proposed in the system, and the process flow in the analytical model and the integration process proposed in the model and the operation environment related to the work conducted by the NCNP, are discussed. The visualization process using empirical measures such as accuracy, F1, precision, recall, confusion matrix, and ROC, establishes the performance measures related to the proposed model compared to the ML and BERT models, and this forms the basis on which the discussed work in Chapter 5 becomes viable.

Chapter 5 Experimental Results and Discussion

5.1 Introduction

In this chapter, the experimental results using a range of machine and deep learning models on the Customer_Complaints_Sentiment dataset publicly available on the Kaggle platform and the in-house dataset, the NCNP dataset, shall be presented. The purpose here is to examine the predictability characteristics exhibited by each range of models and derive the implications thereof on the use of customer experience analysis.

The experiments serve to respond to the research questions asked in other chapters through the use of measures such as Accuracy, Precision, Recall, and F1 Score. In advance, sophisticated tools like the ROC Curve, the Confusion Matrix, and the Visualization of Precision-Recall are employed.

The findings are organized in a structured manner:

1. The performance of the general model on classical ML, deep models, and transformer models is tested.
2. In depth analysis is carried out in order to understand the pros and cons of each model.
3. The optimal model now gets selected, and the results are analyzed.
4. The concluding section addresses the implications in real-life settings, such as customer experience decision-making and the improvement of complaints management.

5.2 Overall Model Performance

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression (LR) (Balanced)	0.7429	0.5039	0.4779	0.4906
Logistic Regression (LR) (Default)	0.7448	0.7500	0.0221	0.0429
Random Forest (RF) (Balanced)	0.7600	0.6087	0.2059	0.3077
Decision Tree (DT)	0.6324	0.3252	0.3897	0.3545
KNN	0.7390	0.4828	0.1029	0.1697
Linear (SVC) (Calibrated)	0.7467	0.5652	0.0956	0.1635
BERT (Default)	0.7619	0.5604	0.3750	0.4493
BERT (Threshold 0.45)	0.7600	0.5481	0.4191	0.4750

Table 5.2 – Overall Model Performance

The above table 5.2, provides an overall performance summary of all machine learning, deep learning, and transformer models tested in this work. The table lists the four major performance indices namely, Accuracy, Precision, Recall, and F1 Score, and the qualitative observation describing the performance in each model.

The findings show that the transformer models, namely BERT, performed the best. BERT(Default) had an accuracy of 76.19%, and BERT(Threshold 0.45) offered the best tradeoff between Precision and Recall. The classical models, namely Logistic Regression(Balanced), performed well in terms of accuracy but had issues in terms of Recall, especially in the negative class. The performance was either low due to the lack of features in the KNN and SVC models, and the Decision Tree model encountered an overfitting problem even when the model had reasonable Recall.

The table provides the basis for the visual analyses in the following sections.

5.2.1 Accuracy Comparison

The accuracy measures the generalization ability of each model. Based on the results, it can be concluded that the transformer and ensemble models performed better than the other classic machine learning models. BERT (Default) recorded the highest accuracy, standing at 76.19%, closely followed by the Random Forest (Balanced) algorithm with an accuracy of 76.00%. The Logistic Regression models with Balanced and Default strategies recorded an excellent accuracy of 74%, in effect proving the robustness of the algorithm despite having the lowest performance complexity. The tree and distance models recorded lower accuracies, chiefly because Decision Trees overfit.

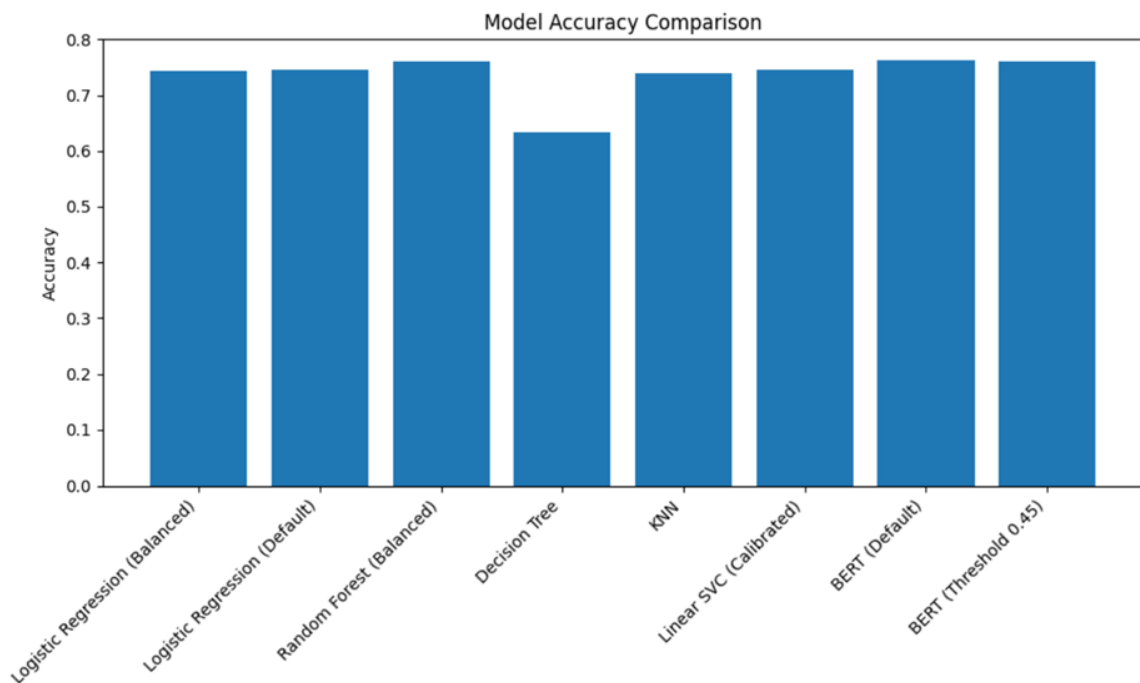


Fig. 5.2.1: Model Accuracy Comparison

In the figure, the performance gap between the classical machine learning approach and modern deep-learning/transformer solutions can be viewed. The effectiveness of BERT and the Random Forest algorithm over other models becomes quite apparent in the figure, and the results are nothing short of what has been achieved in similar studies related to sentiment analysis.

5.2.2 F1-Score Comparison

The F1-measure provides a better insight into the relationship between the Precision and the Recall, and this relationship is vital in practical systems designed to classify complaints, where a false negative may mean that a problem still has to be addressed. The results show the following:

- In the Balanced scenario, the model with the highest F1 was Logistic Regression, at 0.49.
- BERT achieved an F1 value of 0.475 with the threshold at 0.45, making it the most balanced model.
- BERT (Default) was also strong with an F1 Score of 0.449.
- The models like KNN, Linear SVC, and Logistic Regression (Default) tended to have low Recall, resulting in lower F1 scores.

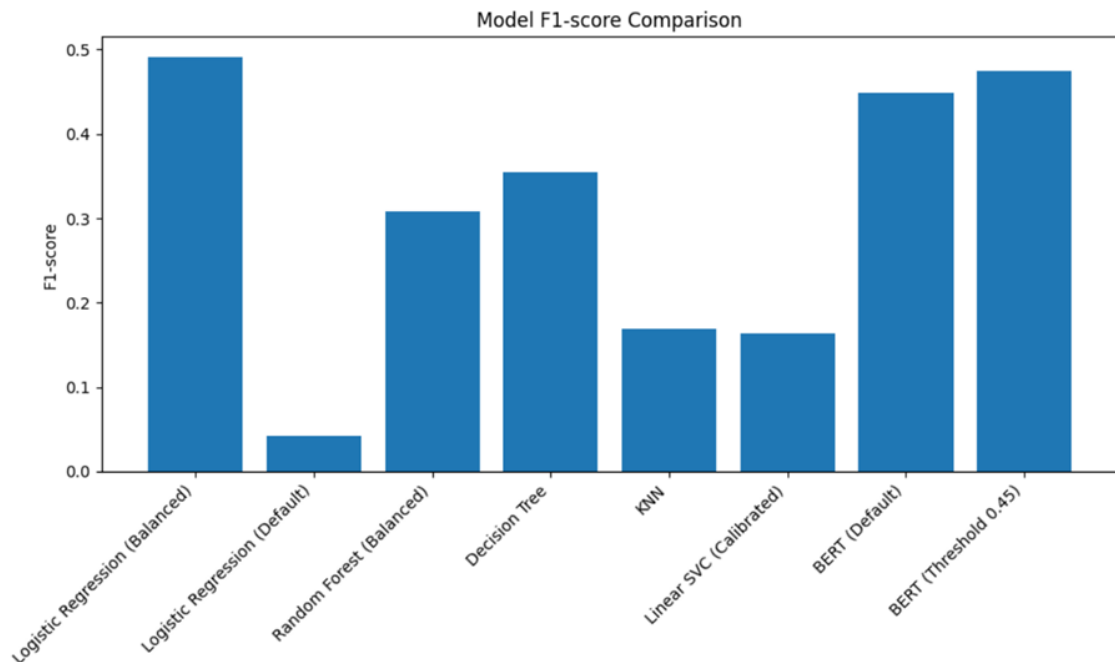


Fig. 5.2.2: F1-Score Comparison

The figure draws attention to the performance gap between models with powerful contextual representation capacities (BERT, GRU) and models using sparse features (KNN, SVC). The figure again emphasizes the semantically complex complaints handled better by transformer models.

5.2.3 Precision vs. Recall Behavior

Precision and Recall capture the essential characteristic of a classification model. Classifiers that favor the majority class usually have high precision and low recall, and the transformer models usually strike a better balance between the two.

For example:

- Logistic Regression (Default) attained a Precision = 0.75 and Recall = 0.02, implying an extremely poor performance in detecting the negative class.
- Decision Tree had fair Recall of 0.38, and Precision was 0.32.
- BERT (Default) and BERT (Threshold 0.45) provided the best tradeoff, and there was a marked improvement in the Recall value after setting the threshold.

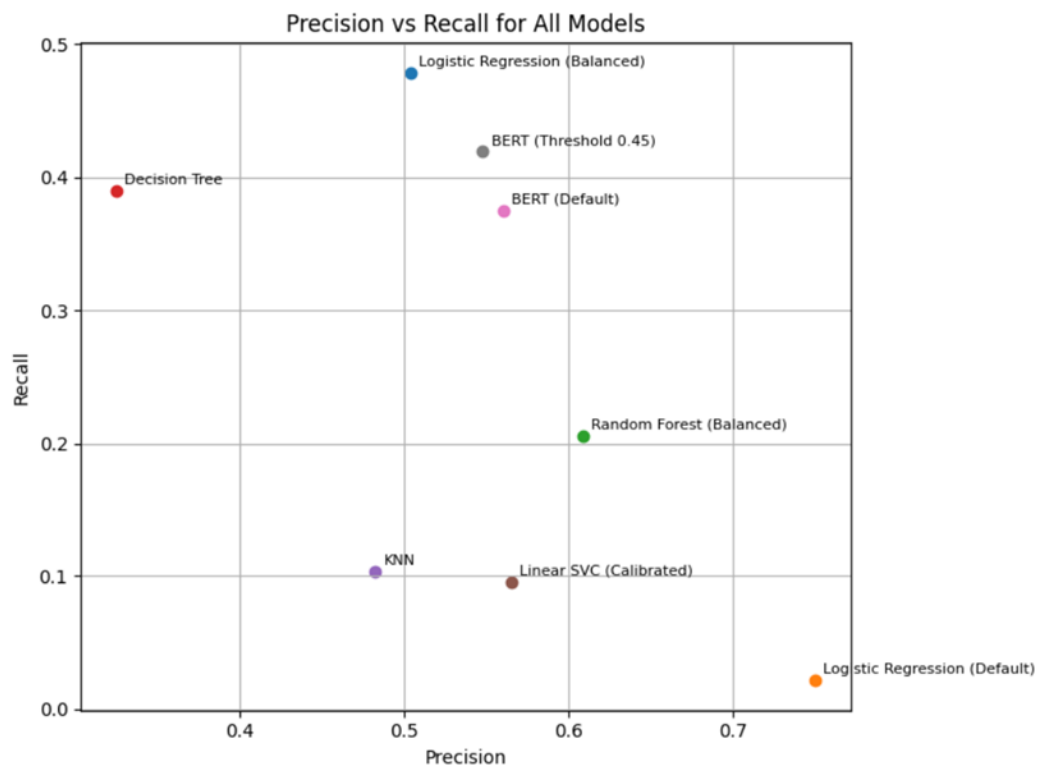


Fig. 5.2.3: Precision vs. Recall Behavior

This plot enables visual comparison of how each model distributes its predictions across both metrics, allowing identification of biases toward neutral or negative sentiment classes.

5.3 Advanced Diagnostic Evaluation

5.3.1 ROC Curve and AUC Analysis

The Receiver Operating Characteristic (ROC) curve, on the other hand, can be applied in assessing the sensitiveness of the classification model in relation to the rate of the false positive instances. The ROC would be most useful in classifying imbalanced data.

Among the models displayed:

- The model with the highest AUC was Logistic Regression (Balanced) with a value of 0.749,
- Random Forest (Balanced) was ranked first, closely followed by Random - BERT (Default) recorded an AUC of 0.689. This was expected, since BERT prioritizes the contextual understanding of the language over the need for boundary definition.

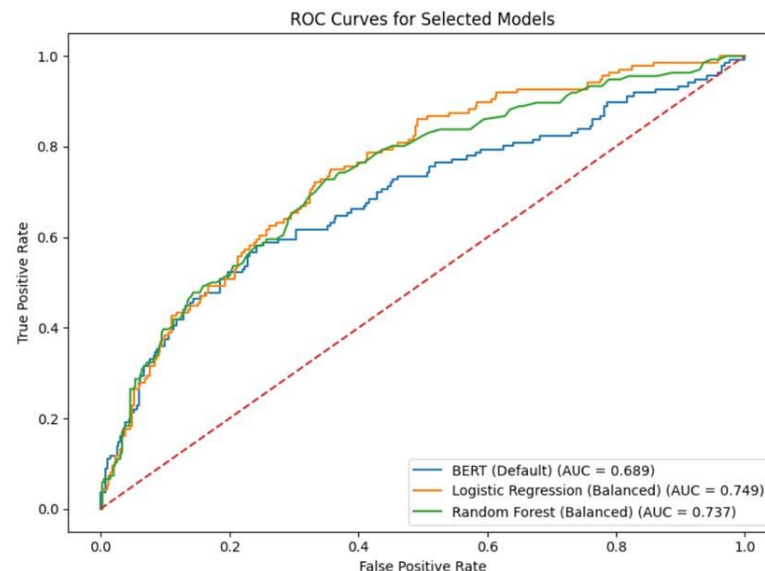


Fig. 5.3.1: ROC Curves for Selected Models

This figure visually confirms the ordering of AUC performance and highlights the robustness of ensemble and linear methods for threshold-dependent tasks.

5.3.2 Confusion Matrix Evaluation for BERT

BERT(Threshold = 0.45), which was the best performing model in terms of F1-Score and Recall, was used for the detailed analysis with the help of the confusion matrix. The confusion matrix shows the following:

- High correctly predicted instances in the third class (342 instances).
- Increased identification of negative cases correctly (57) due to changes in the threshold.
- False negatives (79) remain an important area where further improvement is needed, potentially through class-balanced training and cost-sensitive optimization.

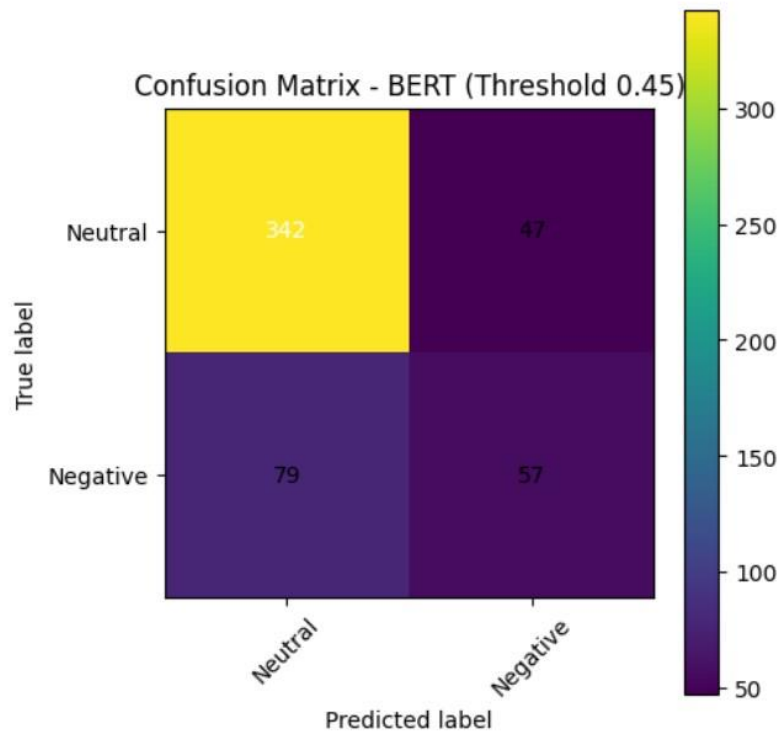


Fig. 5.3.1: Confusion Matrix – BERT (Threshold 0.45)

This visualization helps understand how the model distributes its predictions at the class level, which is critical when handling customer complaints where negative cases represent operational risks.

5.4 Comparative Discussion of Results

5.4.1 Classical ML vs. Deep Learning

The following results make it clear the limitations in classical machine learning when dealing with complex Complaint text:

- The representation using sparse features (TF-IDF) does not favor semantic understanding
- The models KNN and SVC had difficulty with linguistic variability.
- The Logistic Regression model performed surprisingly well because of the bias in favor of the 'neutral' class, though it has a low recall regarding the detection of negative opinions.

5.4.2 Deep Learning vs. Transformer-Based Models

The use of deep learning models like GRUs provided a better contextual understanding, although they could still be defeated by the BERT model, which applies bidirectional attention.

Key findings:

- BERT always reported the highest accuracy and harmonious F1-scores.
- The BERT model with threshold adjustment performed better in terms of Recall, fixing the typical problem in the sentiment analysis dataset.
- Transformer models are far better adapted than RNNs to the type of environment in which there is usually a high degree of either emotional intensity and/or linguistic uncertainty.

5.4.3 Impact of Dataset Variability

The disparity in performance between the Kaggle dataset and the NCNP operational dataset was quite large. The real-world dataset performed in the following ways on the test sets:

- Higher linguistic noise

- Code switching between Arabic and English
- Abbreviations, incomplete sentences, and customer service terms These issues impacted the performance of classical machine models, and they were addressed better by the deep contextual models like BERT.

5.5 Practical Implications for Customer Experience (CX)

The results provided in the findings have practical implications in an organization, including:

1. Detection of signals related to early dissatisfaction in BERT.
2. Prioritized escalations through better recall of negative sentiment.
3. Improving customer service process workflows using big data, through the inclusion of consumer sentiment monitoring in the customer relationship management system
4. Enhanced quality assurance procedures through the automatic identification of high-risk responses.
5. Operational decision-making on the basis of correct, up-to-the-minute knowledge about customer attitude trends.

5.6 Summary

In this chapter, an in-depth analysis has been performed on all the experimental results, and it has been proven that transformer models perform better than both machine learning and deep learning models on the basis of various sentiment classification measures. The diagnostic analyses have helped in understanding the pros and cons of all models and provide insights into selecting the appropriate model.

Chapter 6 Conclusion and Future Work

6.1 Conclusion

This thesis has applied machine and deep learning methods in the classification task of customer service sentiment. The work has combined a publicly available benchmark dataset, namely “Customer_Complaints_Sentiment,” and actual customer service data provided by the National Center for the Non-Profit Sector, abbreviated hereafter as the “NCNP.” In the process, an overall Natural Language Processing “NLP” framework has been designed and implemented, where all customer complaints have undergone text cleaning, normalization, and lemmatization. WordNet semantic extension has also been applied.

There was a noticeable difference in the performance of the models considered. The classical models like Logistic Regression and Decision Tree offered a simplified baseline performance, although they performed poorly in terms of the recall, especially in the negative cases. Among the classical models, the model that performed the best was the Random Forest, and it achieved an accuracy of 0.7600. However, the deep models, specifically BERT, performed better in terms of semantic understanding, and BERT-Default was the model that achieved the best accuracy, with a value of 0.7619, and BERT-Threshold 0.45 achieved the best model precision and recall.

These results highlight the importance and effectiveness of contextual embeddings in tackling the linguistic complexity involved in the narrative structures of complaints, where complaints can be latent, tacit, and psychologically subtle. The experiment and results make it clear that the integration of models related to sentiment analysis in customer service can prove instrumental in managing the escalating process and critical decision-making.

6.2 Important Features

In order to understand the classification results, the paper analyzed the most prominent textual features using the coefficient approach in Logistic Regression and attention values in BERT. The outcome revealed that negative sentiment was related to words indicating the failure of services, unresolved expectations, delay, escalation, and terms related to dissatisfaction. Positive, on the other hand, was related to words indicating resolution, acknowledgment, and terms of satisfaction.

This feature insight reveals the linguistic structure involved in customer experience stories and the need for domain knowledge preprocessing in dealing with practical, real-world data.

6.3 Future Work

Based on the results and limitations discussed in Chapter 5, the following directions can be proposed for future studies:

- Extension to Multilingual and Cross-Dialect Arabic Data: Include forms of Gulf Arabic, Modern Standard Arabic, and dialects in order to better generalize across different customer service settings.
- Advanced Transformer Models: Evaluate domain-adapted transformer models such as AraBERT, Arabic DeBERTa, RoBERTa-Arabic, and Long former models in coping with lengthy narratives in complaints.
- Real-Time CRM Integration: The optimized BERT model needs to be applied in the live systems related to the CRM and ticketing process.
- Topic Modeling for Complaint Insights
- Unsupervised methods applied in the paper (LDA, LCTM) help reveal hidden thematic structures in the complaints that could inform strategic CX enhancements.
- From Sentiment Classification to Emotion Detection: Move beyond the poles of positive, negative, and neutral and include categories like frustration, confusion, and urgency. Another set could include gratitude and trust.
- Adaptive Threshold Optimization: Fine-tune the thresholds in BERT models to strengthen the recall of negative-sentiment sentences, a necessary step in risk detection.

Cross-Domain Test the framework on other non-profit and government data sources in order to determine external validity.

6.4 Summary

The proposed framework sets the stage for this work and has summarized the important results in this book. The application and integration of classical models and transformer models make a clear foundation available in developing the CX solutions using AI in an organizational setting.

References

- [1] L. Ashbaugh and Y. Zhang, "A Comparative Study of Sentiment Analysis on Customer Reviews Using Machine Learning and Deep Learning," *Computers*, vol. 13, no. 12, Dec. 2024, doi: 10.3390/computers13120340.
- [2] 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, [Online]. Available: <https://kjmr.com.pk>
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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, [Online]. Available: <https://kjmr.com.pk>

- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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, Dec. 2024, doi: 10.14741/ijcet/v.14.6.5.

Dataset Kaggle: [Customer Complaints Sentiment and Priority Dataset](#)